UNIVERSITÉ
DE GENÈVE

Language against Communication

A matter of subjectivity

Alexandre Kabbach

Memoir for the PhD in Linguistics

University of Geneva

December 2024

Committee

Christopher Laenzlinger	University of Geneva	President
Isabelle Charnavel	University of Geneva	Co-supervisor
Jacques Moeschler	University of Geneva	Co-supervisor
Aurélie Herbelot	Possible Worlds Research	External examiner
Ryan Nefdt	University of Cape Town	External examiner
Sarah Carvallo	University of Lyon 1	External examiner

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES LETTRES

IMPRIMATUR

DOCTORAT ÈS LETTRES

Linguistique

Thèse de Alexandre KABBACH

Intitulée : « Language against Communication: a matter of subjectivity »

* La Faculté des lettres, sur le préavis d'une commission composée de Mesdames et Messieurs les Professeur-es, Christopher Laenzlinger, président du jury ; Isabelle Charnavel, co-directrice de thèse ; Jacques Moeschler, co-directeur de thèse ; Aurélie Herbelot (Possible Worlds Research); Ryan Nefdt (Université de Cape Town/Université de Bristol) ; Sarah Carvallo (Université Claude Bernard, Lyon 1), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 6 décembre 2024

La Doyenne : Francesca SERRA

Thèse N° 1161

N.B. – La thèse doit porter la déclaration précédente * et remplir les conditions énumérées dans les informations pour la publication d'une thèse.

Un exemplaire numérisé doit être remis à la Direction de l'information scientifique.

Abstract

Do we really understand each other when we speak, sign, or use language in general? This (computational) linguistics thesis begins with a fundamental consideration: we, as individuals, are all singular beings who always make sense of language and the world around us through the privacy of our own minds and the potential idiosyncrasy of our mental content. From there on it asks: how could we possibly understand one another given the singularity and potential incommensurability of our respective subjectivities? It then makes three contributions. First, it argues that mutual understanding is actually a *presupposition* in our theories of language and communication—an assumption that we take from granted rather than one that unfolds naturally from careful empirical observation. Second, it introduces Subjective Coordination Theory—a new proposal for a theory of communication that does not rest on the presupposition of mutual understanding and, as such, dispenses itself from having to posit the existence of shared meaning, shared language and objective communication success in general. Last, it contrasts the concept of subjectivity with that of *normality* so as to better make sense of language models and the field of artificial intelligence at large. This last contribution itself is threefold. First, it argues that language models such as ChatGPT are the product of a particular epistemology called the *epistemology of normalism* which corresponds to a major shift in the philosophy of science in the nineteenth century made possible by the emergence of statistics. Second, it introduces *normal language*—the ontology of language behind language models (i.e. what it is that language models are actually models *of*)—and explains why if language models speak normal language *nobody* actually speaks normal language in practice. Last, it distinguishes *intelligence* from *smartness* to explain precisely why language models such as ChatGPT cannot be expected to ever pass the Turing test. Such models model *ideal* human behavior—ideally error-free—while passing the Turing test specifically requires machines to demonstrate *real* human behavior that deviates from this normative ideal of human behavior. In practice indeed, real people "make mistakes"—they always deviate from whichever normative ideal they live by—and that is also precisely what makes them human. It concludes that such models are therefore models of *artificial smartness* rather than artificial intelligence *per se* and that subjectivity, then characterized as the intrinsic and singular deviation from the norm of our respective individualities, can be understood as yet another fundamental expression of our humanity.

Résumé

Se comprend-on vraiment lorsqu'on parle, signe, ou utilise le langage de manière générale ? Cette thèse de linguistique (computationnelle) débute par une considération fondamentale : en tant qu'individus, nous sommes toutes et tous des êtres singuliers, qui faisons sens du langage et du monde autour de nous dans la confidentialité de nos esprits et à travers les particularités de nos espaces mentaux. Dès lors, elle s'interroge : comment pourrions-nous nous comprendre étant donné la singularité et l'incommensurabilité potentielle de nos subjectivités respectives ? Elle propose ensuite trois contributions. Premièrement, elle défend l'idée que la compréhension mutuelle est en réalité un *présupposé* de nos théories du langage et de la communication ; une considération que l'on prend pour acquise plutôt qu'un énoncé qui découle naturellement d'une étude empirique minutieuse. Deuxièmement, elle présente la *théorie de la coordination subjective* ; une nouvelle théorie du langage et de la communication ne reposant plus sur le présupposé de la compréhension mutuelle et qui, à ce titre, se dispense de la nécessité de postuler l'existence de sens partagé, de langue partagée, ou même d'une caractérisation objective du succès de la communication. Enfin, elle contraste le concept de subjectivité avec celui de *normalité* afin de mieux saisir le fonctionnement des modèles de langue et du domaine de l'intelligence artificielle au sens large. Cette contribution est elle-même triple. Premièrement, elle défend l'idée que les modèles de langue tels que ChatGPT sont le produit d'une épistémologie particulière qu'elle appelle *l'épistémologie du normalisme* et qui correspond à un changement de paradigme majeur au sein de la philosophie des sciences du XIXème siècle rendu possible par l'emergence des statistiques. Deuxièmement, elle présente l'ontologie du langage derrière les modèles de langue (i.e., de quoi ces modèles de langue sont en fait les modèles) qu'elle appelle *langage normal*, et explique pourquoi les modèles de langue parlent le langage normal alors que *personne* ne parle le langage normal en pratique. Enfin, elle distingue le concept d'*intelligence* de celui de *smartness* afin d'expliquer précisément pourquoi on ne peut attendre des modèles de langue tels que ChatGPT qu'ils puissent un jour passer le test de Turing. Elle soutient en effet que ces modèles modélisent un comportement humain *idéal*, idéalement sans erreur, tandis que le test de Turing demande précisément aux machines de pouvoir faire montre d'un comportement humain *réel* qui dévie toujours irrémédiablement de l'idéal normatif qui le sous-tend. Dans la pratique en effet, les êtres humains font toujours des erreurs, et c'est d'ailleurs précisément ce qui les rend humains. Elle conclue donc que ces modèles constituent des modèles de *smartness artificielle* plutôt que d'*intelligence artificielle* au sens propre (de Turing) et que la subjectivité, dès lors comprise comme la déviation intrinsèque de nos individualités respectives par rapport à la norme, peut être interprétée comme une des expressions fondamentales de notre humanité.

Acknowledgments

This PhD has been anything but a lonely journey and I am immensely grateful to all the people who have helped me carry it through.

I am grateful first to Kyoko Hirose Ohara for introducing me to linguistics and without whom I would probably never have started a PhD. I am grateful to Chloé Braud, Sarah Carvallo, Manon Mézière, Anne Reboul, Hannah Rohde, Giuseppe Samo and Lucas Tual for kindly taking the time to answer all my questions regarding computational discourse processing, the history of statistics, the anthropology of gait, the philosophy and evolution of language, the pragmatics of politeness, the cartographic enterprise in syntax and the validity of introspective judgment in linguistics. I am grateful to all the people I have met throughout my various research stays at CIMeC in Rovereto and at the ILLC in Amsterdam—and to Raquel Fernández in particular for her very warm welcome. I am grateful to the people who have helped me make those research stays possible and notably to Frédéric Goubier, Genoveva Puskás and Roxana Vicovanu for helping me draft my doc.mobility proposal for the Swiss National Science Foundation. I am grateful to Eva Capitao, Kristina Gulordava, Luka Nerima, Haozhou Wang and all my colleagues at the University of Geneva for their constant intellectual and emotional support throughout my assistantship, as well as for simple moments of sharing. I am grateful to Caroline & Caroline, Erzsi, Laure, Loraine, Mathilde, Valeria, Yasmine, Zina and the entire ACIL gang at the University of Geneva for the incredible feeling of collective empowerment and for helping me survive this hostile academic world. I am grateful to Corentin Ribeyre for teaching me almost everything I know about computer science, and to Adèle Payet for doing the same with philosophy—and both for so much more than I can actually put into words. I am grateful to Sarah Carvallo, Christopher Laenzlinger and Ryan Nefdt for accepting to be part of my PhD committee. Finally, I am grateful to my co-supervisor Jacques Moeschler for his unfailing availability and for his continuous support throughout this PhD, and to Isabelle Charnavel for accepting to take over upon Jacques's retirement. Without them, none of this would have been possible. Last, but not least, I am grateful to my co-supervisor Aurélie Herbelot for reconciliating me with the concept of mentorship, and for her everyday support which has helped me mature most of the ideas that you will find here written.

Since this thesis is probably as much the product of my own labor than it is of the material conditions that made it possible, I must also acknowledge the financial support of the Italian–Swiss research foundation for a 12 months research grant covering my various research stays at the University of Trento in 2020; the support of the Swiss National Science Foundation for an 18 months doc.mobility research grant covering my stay at the University of Amsterdam in 2021–2022; the support of the Ernst and Lucie Schmidheiny foundation for a dissertation writing grant covering a 6 months period in 2023; and the support of the French unemployment benefit agency Pôle Emploi (now France Travail) for the very last months of my PhD.

Table of Contents

Table of Contents v

List of Algorithms ix

List of Axioms x

List of Definitions xi

List of Figures xiii

List of Tables xvi

Introduction 1

Chapter 1 Language for communication 7

1.1 Summary 7
1.2 Communication and the problem of ambiguity 8
  1.2.1 The code model of communication 8
  1.2.2 Ambiguity against the communicative function of language 11
1.3 Ambiguity against communication: an artificial problem? 13
  1.3.1 Ambiguity against ambiguous systems 13
  1.3.2 Putting linguistic ambiguity back into communicative context 14
  1.3.3 Making sense is a process 19
1.4 Is language an effective communication system? 23
  1.4.1 Underdetermination as ambiguity applied to communication 23
  1.4.2 Subjectivity against effective communication 27
  1.4.3 Why communication is not signaling 33
1.5 Communication beyond mutual understanding 37

1.5.1 The presupposition of mutual understanding . . . 37
1.5.2 The requirement of shared meaning . . . 39
1.5.3 The subjectivist approach to communication . . . 44

Chapter 2 Communication as subjective coordination 49

2.1 Summary . . . 49
2.2 Subjective Coordination Theory . . . 50
2.2.1 A non-normative subjectivism . . . 50
2.2.2 The five axioms of Subjective Coordination Theory . . . 53
2.2.3 Operationalizing the axioms: vectors and projections . . . 57
2.2.4 A theory of acommensurability in language and communication . . . 62
2.3 Subjective coordination in action . . . 65
2.3.1 Intuitions of communication success . . . 65
2.3.2 Dynamics of communication: a computational simulation . . . 70
2.3.2.1 Computational model . . . 70
2.3.2.2 Hypothesis . . . 76
2.3.2.3 Experimental setup . . . 77
2.3.2.4 Results . . . 78
2.3.3 From subjective coordination to collective intelligence . . . 85
2.4 Language beyond communication . . . 91
2.4.1 Function, evolution and adaptation . . . 91
2.4.2 Towards a non-theological evolutionary linguistics . . . 96
2.5 Intermediate conclusion . . . 100

Chapter 3 Subjectivity against normality 103

3.1 Summary . . . 103
3.2 Deconstructing artificial intelligence . . . 104
3.2.1 Making sense of ChatGPT . . . 104
3.2.1.1 Language, models and understanding . . . 104
3.2.1.2 From GPT-2 to ChatGPT . . . 108
3.2.1.3 From correct behavior to normal language . . . 111
3.2.1.4 From human intelligence to artificial smartness . . . 112
3.2.1.5 ChatGPT: a smart model of normal language . . . 113
3.2.2 Norm, normal, normative: a history of normalism . . . 114
3.2.2.1 Artificial intelligence, or statistics on steroids . . . 114
3.2.2.2 Normality, biology, and statistics . . . 116
3.2.2.3 Normality and normativity in language and intelligence . . . 120
3.2.2.4 The epistemology of normalism . . . 122
3.3 From correct behavior to artificial smartness . . . 127
3.3.1 Correct behavior . . . 127
3.3.1.1 The correctness principle . . . 127
3.3.1.2 Should machines really “think” like humans? . . . 129
3.3.1.3 The possible and the probable . . . 133

3.3.2 Normal language 136
  3.3.2.1 Normal language as social language 138
  3.3.2.2 Normal language as normative language 142
  3.3.2.3 Normal language as objective language 146
  3.3.2.4 Normal language as neutral language 150
  3.3.2.5 Normal language as single language 154
  3.3.2.6 Normal language as shared language 161
3.3.3 Smart machines 167
  3.3.3.1 Normal against exceptional, or intelligence against smartness 172
  3.3.3.2 From the normal as familiar to the statistically normal . . 182
  3.3.3.3 Intelligence: a scientific concept? 190
3.4 Towards a subjectivist science of human being 192
3.4.1 Towards a science of structure without order 193
  3.4.1.1 Machine bipedalism 193
  3.4.1.2 Beyond normality and internal order 196
  3.4.1.3 Beyond human nature and external order 199
3.4.2 Towards a subjectivist linguistics 203
  3.4.2.1 Bringing back subjectivity 203
  3.4.2.2 Subjectivity is not noise 205
  3.4.2.3 From subjectivity to subjectivism 207

Conclusion **213**

Appendix A ChatGPT **217**
A.1 ChatGPT and the Robinson experiment 218
A.2 ChatGPT and the lexicon 220
  A.2.1 Gobbledygook 220
  A.2.2 Ulotrichous 221
  A.2.3 Kakorrhaphiophobia 222
A.3 ChatGPT and spelling 223

Bibliography **229**

List of Algorithms

1 Making sense . . . 72
2 Interpretation . . . 73
3 Reasoning . . . 74
4 Memorizing . . . 74
5 Projection . . . 74
6 Distance . . . 75

List of Axioms

1 Productivity 54
2 Atomic Compositionality 54
3 Underdetermination 54
4 Making Sense 54
5 Least Effort 54

List of Definitions

1 Code 10
2 Ambiguous communication system 14
3 Utterance 15
4 Communication as intention resolution 24
5 Cooperative principle 25
6 Maxims of conversation 25
7 Cognitive principle of relevance 28
8 Communicative principle of relevance 29
9 Presumption of optimal relevance 29
10 Relevance-guided comprehension heuristic 29
11 The semantic problem 33
12 The effectiveness problem 33
13 Dogma of the double conformity 40

14 Vector space 58
15 Metric 59
16 Basis 60
17 Projection 61
18 Principle of intersubjectivity 65
19 Principle of homogeneity 66
20 Virtuous circle of mutual understanding 67
21 Virtuous circle of communication 70
22 Conceptual homogeneity 71
23 Communicative cost 71
24 Intersubjective coordination hypothesis 77
25 Subjective coordination hypothesis 77
26 Adaptation 93

27 Reinforcement learning . . . 109
28 Aphasia . . . 117
29 Normalism . . . 122
30 Humanism . . . 123
31 Order relation . . . 126
32 Correctness principle . . . 127
33 Normal language . . . 136
34 Language model . . . 137
35 Perplexity . . . 137
36 Social realism . . . 147
37 Statistical realism . . . 148
38 Dogma of sentence meaning . . . 165
39 Dogma of saying . . . 165
40 Intelligence . . . 168
41 Smartness . . . 168

List of Figures

1.1 The original communication diagram under the code model of Shannon & Weaver (1949/1964). 9
1.2 Linguistic communication as the encoding and decoding of messages. 10
1.3 Semantics of the *shape* feature in French road signs. 13
1.4 Ambiguity of the *shape* feature in French road signs. 13
1.5 I saw the man [with the telescope] 17
1.6 I saw [the man with the telescope] 17
1.7 Noise correction in the code model of communication. Original diagram from (Shannon & Weaver, 1949/1964, Figure 8, p.68). 32

2.1 A three-dimensional representation of four word vectors in a semantic space, with their corresponding coordinates (borrowed from Lenci, 2018, Figure 1). . . . 60
2.2 An example of three different parallelepipeds with different sets of coordinates, all corresponding to the exact same black parallelogram when projected on the xy plane along the z axis. 61
2.3 Communicative interaction between three agents (a_1, a_2, a_3). Arrows show all possible options for agents to sample a projected concept (i.e. communicated signal) from each other. Each agent can only sample a single concept per iteration: a_1 must choose between l_{21} and l_{31}, for instance. 72
2.4 Average communicative cost across two agents, per iteration, *with* and *without* smoothing, in an intersubjective coordination scenario (no projection) with near-null initial homogeneity: $N_a = 2$, $N_d = 0$, $H_i \sim 0$. 79
2.5 Average communicative cost across two agents ($cost$) and homogeneity ($homog$), per iteration, in an intersubjective coordination scenario (no projection) with near-null initial homogeneity: $N_a = 2$, $N_d = 0$, $H_i \sim 0$. 80

2.6 Average communicative cost across N_a agents, per iteration, in an intersubjective coordination scenario (no projection) with near-null initial homogeneity: $N_a \in [2, 5, 10]$, $N_d = 0$, $H_i \sim 0$. 81
2.7 Homogeneity across N_a agents, per iteration, in an intersubjective coordination scenario (no projection) with near-null initial homogeneity: $N_a \in [2, 5, 10]$, $N_d = 0$, $H_i \sim 0$. 81
2.8 Average communicative cost across two agents (*cost*) and homogeneity (*homog*), per iteration, in a subjective coordination scenario with one projected dimension and near-null initial homogeneity: $N_a = 2$, $N_d = 1$, $H_i \sim 0$. 81
2.9 Average communicative cost across two agents, per iteration, in both intersubjective (no projection) and subjective coordination scenarios, with N_d projected dimensions and near-null initial homogeneity: $N_a = 2$, $N_d \in [0, 1, 2, 5]$, $H_i \sim 0$. . 82
2.10 Homogeneity across two agents, per iteration, in both intersubjective (no projection) and subjective coordination scenarios, with N_d projected dimensions and near-null initial homogeneity: $N_a = 2$, $N_d \in [0, 1, 2, 5]$, $H_i \sim 0$. 82
2.11 Average communicative cost across two agents, per iteration, in a subjective coordination scenario with one projected dimension and H_i initial homogeneity: $N_a = 2$, $N_d = 1$, $H_i \in [0, .75, .85, .95]$. 84
2.12 Homogeneity across two agents, per iteration, in a subjective coordination scenario with one projected dimension and H_i initial homogeneity: $N_a = 2$, $N_d = 1$, $H_i \in [0, .75, .85, .95]$. 84

3.1 The normal curve, centered on the mean/average (dashed line). 118
3.2 The *cumulative* normal curve, centered on the mean/average (dashed line). . . 121
3.3 Artificial intelligence, or the modeling of *normal* human behavior. Correct machine behavior corresponds to normal/average human behavior marked as a straight blue line at the center of a normal curve modeling a hypothetical distribution over human behavior. 178
3.4 Artificial smartness, or the modeling of *exceptional* human behavior. Correct machine behavior corresponds to exceptional human behavior marked as a straight blue line at the far right of a cumulative normal curve modeling a hypothetical distribution over human behavior. Note how the straight blue line marking exceptional human behavior specifically deviates from the normal/average dashed line at the center of the curve. 178

3.5 Why language models such as ChatGPT are unlikely to pass the Turing test: they are models of artificial smartness where progress is driven by the straight blue line at the extreme far right of the cumulative normal curve (i.e. exceptional human behavior) while passing the Turing test requires targeting the dashed line at the center of the cumulative normal curve (i.e. normal human behavior). Machine behavior is here categorized into three zones: the left red dashed zone where it is too bad to be human, the center green dotted zone where it is good enough to be human, and the right red dashed zone where it is too good to be human. Segmentation between those three zones is completely arbitrary and chosen for illustrative purposes only. 179
3.6 Comparative illustration of skeletons of primates, from (Huxley, 1863/2009). . . 194

A.1 Replication of the Robinson experiment of (Bender & Koller, 2020) with ChatGPT v3.5, accessed August 23 2023. [1/2] 218
A.2 Replication of the Robinson experiment of (Bender & Koller, 2020) with ChatGPT v3.5, accessed August 23 2023. [2/2] 219
A.3 Definition of “gobbledygook” provided by ChatGPT v3.5, accessed March 10 2024. 220
A.4 Definition of “gobbledygook” provided by the Merriam–Webster online dictionary, accessed March 10 2024. 220
A.5 Definition of “ulotrichous” provided by ChatGPT v3.5, accessed March 10 2024. 221
A.6 Definition of “ulotrichous” provided by the Merriam–Webster online dictionary, accessed March 10 2024. 221
A.7 Definition of “kakorrhaphiophobia” provided by ChatGPT v3.5, accessed March 10 2024. 222
A.8 Definition of “kakorrhaphiophobia” provided by the Merriam–Webster online dictionary, accessed March 10 2024. 222
A.9 Replication of the Robinson experiment of (Bender & Koller, 2020) with ChatGPT v3.5, accessed March 10 2024, but with the explicit instruction to use imperfect spelling. 223
A.10 Instruction given to ChatGPT v3.5 to rewrite the output it provided as a response to the Robinson replication experiment in Figure A.1 and Figure A.2 by adding spelling mistakes. [1/2] 224
A.11 Instruction given to ChatGPT v3.5 to rewrite the output it provided as a response to the Robinson replication experiment in Figure A.1 and Figure A.2 by adding spelling mistakes. [2/2] 225
A.12 Response from ChatGPT v3.5, accessed March 10 2024, to the instruction provided in Figure A.10 and Figure A.11. [1/2] 226
A.13 Response from ChatGPT v3.5, accessed March 10 2024, to the instruction provided in Figure A.10 and Figure A.11. [2/2] 227

List of Tables

1.1 Why communicative effectiveness is neither *necessary* nor *sufficient* to attest of mutual understanding (effectiveness $\not\Leftrightarrow$ mutual understanding). 36

2.1 List of hyperparameters and their corresponding values. 78

3.1 Intelligent behavior against smart behavior. 113
3.2 Artificial intelligence against artificial smartness. 122
3.3 Intelligence against smartness, or normality against exceptionality. 170
3.4 Recap of Turing test score (defined as the probability of correct machine identification by a normal/average human interrogator) in idealized scenarios, considering the human population to be split into two complementary sets (W and $\overline{\mathrm{W}}$) with two fully distinct and incompatible normative ideals of normal behavior that cannot be evaluated beyond chance by members of the other set. The last two configurations (full row of W and full row of $\overline{\mathrm{W}}$) correspond to situations where the test would provide meaningful test scores but would only evaluate normal W behavior and normal $\overline{\mathrm{W}}$ behavior rather than normal *human* behavior. 189

Introduction

This thesis actually started as a computational linguistics thesis. Like any "regular" computational linguistics thesis, it would begin with a *task*—a particular "problem" the machine would have to solve—paired with an *evaluation*—a quantitative process by means of which machine performance on the task at hand would be assigned a particular *score*. My goal would then be to design new computational models so as to improve those scores in comparison to the state-of-the-art (Sparck Jones & Galliers, 1995; Paroubek et al., 2007; Church, 2017, 2018; Church & Hestness, 2019).

For me, that task was *lexical similarity* (Rubenstein & Goodenough, 1965; Miller & Charles, 1991; Bruni et al., 2014; Hill et al., 2015; Gerz et al., 2016). My specific goal was to create computational models of word meaning that would align best with human judgment—considering, for instance, that *cat* is more similar to *dog* than to *sofa*. To do so, I relied on *distributional semantic models* (Lenci, 2018; Lenci et al., 2022) which operationalize the *distributional hypothesis* of Harris (1954) and Firth (1957/1962) telling us that "difference of meaning correlates with difference of [context] distribution" (Harris, 1954, p.156) and that "you shall know a word by the company it keeps" (Firth, 1957/1962, p.11).[1] Its basic idea is that a correlation exists between *distributional* and *semantic* similarity, so that the former can be used to estimate the latter (Sahlgren, 2008, p.34). Concretely, distributional semantic models generate vector representations of lexical items by aggregating (e.g. Turney & Pantel, 2010) or predicting (e.g. Mikolov et al., 2013) co-occurrence patterns of word collocates from corpus data. Their geometric structure then makes it possible to compute distances between vectors and output measures of semantic similarity between their corresponding lexical items—the "closer" two lexical vectors are, the more "similar" the meaning of the words they model should be.

Rapidly though, this task started bugging me. How could we possibly compute the

---

[1] An idea that is often paraphrased as "words which are similar in meaning occur in similar contexts" (Rubenstein & Goodenough, 1965, p.627) or as "words that occur in the same contexts tend to have similar meanings" (Pantel, 2005, p.126).

"absolute" similarity between words as if it was some sort of "objective" matter, as if words were always intrinsically similar or dissimilar, or as if words themselves always had a single, clear, context-independent and speaker-independent meaning, when everything about my experience of language actually suggested otherwise? The field of computational linguistics is aware of this tension, of course—which only made my frustration grow stronger. Faruqui et al. (2016, p.31), for instance, call the very notion of lexical similarity "subjective" while Batchkarov et al. (2016, p.8) tell us that "[w]ords are [. . .] not inherently similar or dissimilar". Harris (1954, p.152) himself—the "father" of distributional semantics—argues that "[w]e cannot say that each morpheme or word has a single or central meaning, or even that it has a continuous or coherent range of meanings". Confronted with those contradictions, I thus felt the need to get back to basics so as to better understand what I was actually doing. This manuscript is an account of this intellectual journey.

Getting back to basics meant for me starting from very general questions—what is language? What is its purpose? Its function?—as the underlying assumptions of computational linguistics always seemed to bring me back to the communicative function of language one way or another. What I soon found out, however, was that this communicative function of language was far from being universally endorsed by the literature (e.g. Chomsky, 2016). This observation gave rise to the fundamental opposition that you find in the title of this thesis: *language against communication*. This was the opposition I needed to start from, and ultimately overcome. Yet, while for most of the literature the opposition usually boils down to the question of the proper characterization of the notion of "function"—and to that, more specifically, of whether language could be said to have evolved *for* the purpose of communication (e.g. Origgi & Sperber, 2000; Reboul, 2017)—for me it meant something different. It meant having to question the characterization of "communication" as the mechanism by means of which we can convey our thoughts to others, and the characterization of "language" as the pairing of sound (or sign) with (objective) meaning.

My problem was this: how could we possibly understand one another in communication given that we always made sense of language and the world around us through the privacy of our own minds and the potential idiosyncrasy of our mental content? How could we objectivize a shared language with shared meaning when everything about my experience of language and meaning suggested that those were but a matter of *subjectivity*; that we could always use language in our own ways, hold different beliefs and attach different concepts to the same words, and ultimately ascribe different meanings to the same linguistic expressions? To me, objectivizing language and meaning always seemed to imply having to marginalize subjectivity one way or another—either by saying that it did not exist (that our singularities were not so significant after all) or that it did not really matter (that mutual understanding could proceed regardless)—and I just couldn't bring myself to it. I couldn't bring myself to it as it always seemed to create irreconcilable tensions with my own experience of language and communication. I wanted to reconcile theory with experience—with *my* experience, at least—so I decided to reconstruct a theory of language and communication starting from my subjectivity and with the clear intention to design a theory that would no longer function *despite* or even *against* subjectivity but *around* it this time. There, I had my research question:

> How can we construct a theory of language, communication and human being in general, that does not marginalize subjectivity and the singularity of individuals?

This thesis is a contribution towards this goal.

Its contribution is threefold. Chapter 1 argues that mutual understanding is a *presupposition* in our standard theories of language and communication—an assumption that we take from granted rather than one that unfolds naturally from careful empirical observation. Chapter 2 introduces Subjective Coordination Theory: my proposal for a psychological theory of communication that does not rest on the presupposition of mutual understanding and formalizes communication as the *subjective* (rather than objective, or even intersubjective) coordination of minds. Chapter 3 proposes to approach subjectivity as our intrinsic and singular deviation from the *norm* at the individual level, and as what makes us all fundamentally human in comparison to machines which, themselves, are always designed to strictly follow the normative ideal of human behavior we live by.

Before I turn to the contribution of each chapter in more details, let me clarify a bit more what I mean by "subjectivity". As you will have understood, I use the term more or less synonymously with "singularity" or "individual variability". Psychologically speaking, I use subjectivity to refer to the privacy of the mind and the potential idiosyncrasy of mental content—which does not necessarily imply any form of consciousness or self-awareness of the "subject" whatsoever. Biologically speaking, I use subjectivity to refer to the singularity of a living organism in general, both at the genetic level and the phenotypic level. Sociologically speaking, I use subjectivity to refer to the singularity of social experience, and to the fact that we, as individuals, always characterize singular intersections of otherwise shared collective experiences. Developmentally speaking, I use subjectivity to refer to the singularity of our environments and environmental experiences. And phenomenologically speaking, I use subjectivity to refer to the fact that we, as individuals, always make sense of language and the world around us in our own ways, within the privacy of our own minds and the singularity of our respective background experiences.

Coming to terms with the idea that mutual understanding constitutes a presupposition in our standard theories of language and communication has not been an easy task, which is precisely why I dedicate Chapter 1 to the detailing of that thought process. I will tell the story as it unfolded for me, introducing questions and arguments as they came to me as I went along. I will start from the question of the communicative function of language, and more specifically from the controversy surrounding the problem posed by ambiguity for communicative *efficiency*. I will deconstruct the argument of the communicative efficiency of language and show that it presupposes communicative *effectiveness* and through it, the possibility of mutual understanding. I will then question the very possibility of mutual understanding in light of the challenge posed by subjectivity, which will allow me to conclude that it actually constitutes an “axiom” in our theory of language and communication, and an axiom that could and should be changed.

Crucial to the understanding of my contribution in this thesis is that it does not try and settle on the question of *whether* we understand one another in communication but rather asks *why* we need to posit that we can do so within our theories of language and communication. Crucial to the understanding of my concept of “mutual understanding” then is that it is not restricted to the strict identicity of interpretations (or messages) across interlocutors but that it encompasses the relative similarity (or partial overlap) between those interpretations as well. What characterizes mutual understanding throughout this thesis indeed is that it formalizes communication success in *objective* (and normative) terms—and thus fundamentally requires us to *compare* interpretations across interlocutors to assess communication success. In Subjective Coordination Theory, however, interpretations *need not* be compared across interlocutors as there are no “correct” or “incorrect” interpretations—and no “more” or “less” correct interpretation either. This does not mean that we can no longer talk about communication success, however, but only that “success” is now framed in purely *subjective* (and non-normative) terms—and more specifically, as the ability for an interpreter to find or construct an interpretation of a communicated signal compatible with the potential singularity of their respective conceptual space.

Now, it is one thing to argue that a certain proposition constitutes an “axiom” in a given theory, and yet another to argue that it should be changed. So why should we abandon the presupposition of mutual understanding? My goal in Chapter 2 will be to show that doing so can lead to *simpler* (if not *better*) theories of language and communication—“better” to the extent that they can prove equally capable of accounting for some of our crucial intuitions about language and communication but can do so in “simpler” terms insofar as they no longer have to compensate (through various theoretical mechanisms) for the problems posed by the marginalization of subjectivity. Doing so, however, requires understanding first what it is that our intuitions of communication success are supposed to be intuitions *of* if they are not intuitions of mutual understanding, which is why I will begin by arguing that our intuitions of communication success are actually intuitions of the *smoothness* of our communicative interactions—that is, of how “easy” or “difficult” it is for us to make sense of one another and of our respective communicated signals. I will then try and demonstrate the explanatory power of Subjective Coordination Theory for what I call the *virtuous circle of communication*—the intuition we have that the more we communicate with one another, the smoother our communicative interactions become. I will do this with what I know best: *computational simulations*, but with no claim at cognitive plausibility whatsoever and with the sole intention to use them to demonstrate the absence of a specific (theoretical) necessity. In other words, I will not use those computational simulations to demonstrate what *is* but only to illustrate what *could be.* In effect, my purpose will be to show that the presupposition of mutual understanding is not necessary to account for the virtuous circle of communication and that this can be clearly illustrated by a computational implementation of Subjective Coordination Theory modeling the dynamics of our communicative interactions. I will then conclude this chapter by returning to the question of the *function* of language and discuss the potential contribution of Subjective Coordination Theory for the field of evolutionary linguistics at large.

In Chapter 3, I will return to the question of artificial intelligence with which I originally started this PhD to contrast the concept of subjectivity with that of *normality*. My main contribution in this chapter will be to argue that language models are the product of a particular epistemology which I call the *epistemology of normalism* and which corresponds to a major shift in the philosophy of science in the nineteenth century made possible by the emergence of statistics—a shift which, according to Hacking (1990), replaced the study of "human nature" with that of "normal people". Since it is difficult to write a contribution to the field of artificial intelligence in 2023/2024 without addressing the giant elephant in the room that is ChatGPT, I will begin with a case study of that model which I will characterize as a *smart model of normal language*—in short: a model of how people *ought* to behave rather than how they *do* behave in practice. I will argue that computational models such as ChatGPT model a normative *ideal* of correct and exceptional human behavior that no *real* human actually abides to in practice. In practice indeed, real people make mistakes—they always deviate from the normative ideal of correct and exceptional human behavior they live by—and that is also precisely what makes them human. This explains why, upon detailing the ontology of *normal language* on which they rest, I will argue that *language models speak normal language* and yet *nobody speaks normal language* and why, upon considering their ability to constitute true models of artificial intelligence in the sense of (Turing, 1950), I will argue that such models are unlikely to ever pass the Turing test as the Turing test specifically requires machines to make mistakes and to deviate from the normative ideal of correct and exceptional human behavior those models are usually designed to follow. I will conclude that such models are actually models of what I call *artificial smartness* rather than artificial intelligence *per se*.

This last chapter will highlight a certain number of critical considerations. First, that normal language largely predates language models, as it is indeed the very same ontology of language that we find behind public dictionaries which, as such, can be characterized as the "theoretical ancestors" of language models. Second, that Turing's conception of intelligence remains fundamentally at odds with our every use of the term (or with that of most of the psychological and cognitive science literature for that matter) as for Turing indeed intelligence is not something one can have more or less of but a fundamental ability that one possesses by virtue of being human. Last, but not least, that our practice of building artificial minds is almost systematically driven by a certain "reflex" to align what is correct for machines with what is correct for humans—and yet, this is precisely what the Turing test is asking us *not* to do. All such considerations emphasize what remains, first and foremost, the main contribution of this thesis: a critical reflection of my own *praxis* as a computer scientist and a computational linguist.

CHAPTER

1

Language for communication

1.1 Summary

In this chapter, I propose to tackle the question of the communicative function of language by approaching it through the problem of *ambiguity*. I first define communication as the process by which two or more interlocutors exchange messages through a potentially noisy channel by converting them to transmitted signals by means of a shared code (§1.2.1). I then introduce two opposing views regarding the role played by ambiguity in the communicative function of language: one which argues that ambiguity makes language an *unoptimized* communication system; and the other which argues, on the contrary, that ambiguity makes language an *efficient* communication system (§1.2.2).

I contribute to this discussion by distinguishing *ambiguity* from *ambiguous system* in order to show that the presence of ambiguity alone amongst the features of communicated signals does not suffice to make language an ambiguous *system* as a whole (§1.3.1). I then argue that linguistic ambiguity constitutes an artificial problem from the point of view of communication, in that it always corresponds to artificial cases which arbitrary divorce otherwise unambiguous communicated signals from their communicative context for the needs of linguistic inquiry (§1.3.2). I conclude that language is not an ambiguous system in the "strict" sense, in that ambiguity never prevents the interpretation of *at least one* message by the interpreter(s), but that the crucial question is whether this interpreted message actually corresponds to the original message encoded by the speaker—and so whether language is, in fact, an *effective* communication system in the first place (§1.3.3).

I argue that the structural underdetermination of communicated messages constitutes a major obstacle to the effectiveness of communication, for it implies that every communicated signal always leaves open a multiplicity of options as to how it should be interpreted in practice (§1.4.1). I detail how the subjective nature of the interpretation process definitely challenges the systematic effectiveness of linguistic communication, insofar as interlocutors always make sense of each other's utterances in their own private and potentially idiosyncratic mental spaces (§1.4.2). I distinguish the notion of *mutual understanding* from that of *communicative effectiveness* and show that mutual understanding proves neither necessary nor sufficient to attest of communicative effectiveness (§1.4.3).

I go on to detail how that notion of mutual understanding actually constitutes a *presupposition* in standard theories of communication, that is, an assumption that is taken for granted rather than one that unfolds naturally from careful empirical observation (§1.5.1). From this presupposition of mutual understanding derives what I call the *assumption of shared meaning* which requires that we do share meaning in order to understand one another in communication (§1.5.2). I argue that those assumptions are unwarranted, and conclude on the perspective offered by getting rid of the presupposition of mutual understanding, that would see us approach communication as a *subjective*—rather than *objective*—coordination of minds (§1.5.3).

1.2 Communication and the problem of ambiguity

1.2.1 The code model of communication

The most straightforward way to approach communication is probably to start from the *code model* of Shannon & Weaver (1949/1964), since its conceptual underpinning—namely, that communication consists in the exchange of messages converted to transmitted signals by means of a shared code between interlocutors—is what grounds most theories of communication from Aristotle to modern semiotics (see Sperber & Wilson, 1986/1995, §1.1 for a detailed version of this argument).

The code model intends to provide a very general theory of communication—as attested notably by the change of title from "*A* mathematical theory of communication" in the original (Shannon, 1948) paper to "*The* mathematical theory of communication" in the later (Shannon & Weaver, 1949/1964) book. Under its most general characterization, communication is there defined as the set of procedures by means of which one "mechanism" affects another—and thus as something potentially not limited to verbal or even biological forms. In the context of *human communication*, however, the definition is slightly more restrictive: the mechanism in question is "the mind", although communication—covering "all human behavior"—should apply equally well to music, painting, theatre, or ballet than to speech or text (see Shannon & Weaver, 1949/1964, p.3).

Figure 1.1 reproduces the original diagram illustrating the functioning of communication under the code model (see Shannon & Weaver, 1949/1964, p.7). A transmitter converts a given message (chosen from an information source) to a signal sent to a receiver through a potentially noisy channel (think of noise as interference in a radio channel, for instance). The receiver then converts the received signal back into a message and hands it over to its destination. In the particular case of *verbal communication*, for instance, each component can be interpreted as follows:

> When I talk to you, my brain is the information source, yours the destination; my vocal system is the transmitter, and your ear and the associated eighth

> nerve is the receiver.
>
> Shannon & Weaver (1949/1964, p.7)

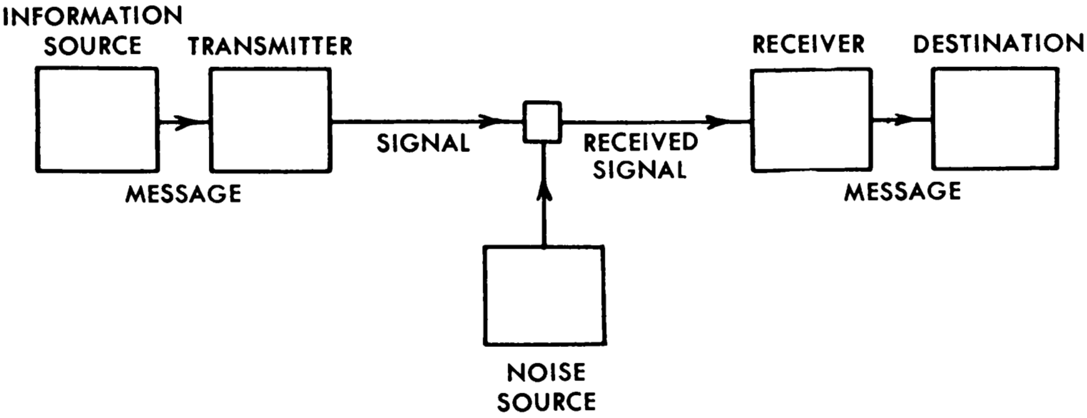


Figure 1.1: The original communication diagram under the code model of Shannon & Weaver (1949/1964).

Crucial to the understanding of the code model then is the notion of "information": communication is initiated by choosing a particular message from an information source and the notion of information serves to quantify the range of possibilities in choosing a particular message in a given situation. In that sense, the technical notion of *Shannon information* is to be distinguished from the more everyday sense of the word "information", notably in that it is not the messages themselves that carry Shannon information but the overall probability distribution over those messages:

> Note that it is misleading (although often convenient) to say that one or the other message conveys unit information. The concept of information applies not to the individual messages (as the concept of meaning would), but rather to the situation as a whole, the unit information indicating that in this situation one has an amount of freedom of choice, in selecting a message, which it is convenient to regard as a standard or unit amount.
>
> Shannon & Weaver (1949/1964, p.9)

The general idea behind this notion of Shannon information is that the less you can predict which message will be chosen next by a given information source, the more informative that source will be.[2]

Be that as it may, the code model is more often than not used simply as a *metaphor*—which Reddy (1979/1993) calls the "conduit" or "containment" metaphor—and so as model of *information transfer* in the more plain sense of the term. As far as linguistic communication

---

[2] Throughout this chapter, I will use the word *information* to refer to the everyday sense of the term and *Shannon information* to refer to the technical sense of the term as per Shannon & Weaver (1949/1964).

is concerned, the overall intuition is that language provides the means by which human beings can convey their thoughts to others—an idea which we find explicitly spelled out in (Locke, 1690/2004, especially book III, ch.1 and ch.2), for instance:

> Besides articulate sounds therefore, it was further necessary, that he [human being] should be *able to use these sounds, as signs of internal conceptions*; and to make them stand as marks for the ideas within his own mind, whereby they might be made known to others, and the thoughts of men's minds be conveyed from one to another.
>
> Locke (1690/2004, Book III, ch.1 §2)

Human linguistic communication can thus be understood as a specific type of interaction taking place between at least two interlocutors—a *speaker* and an *interpreter*—who encode and decode messages transmitted through a potentially noisy channel (Figure 1.2).[3] The notion of "code" can then be given a more precise characterization (Definition 1).

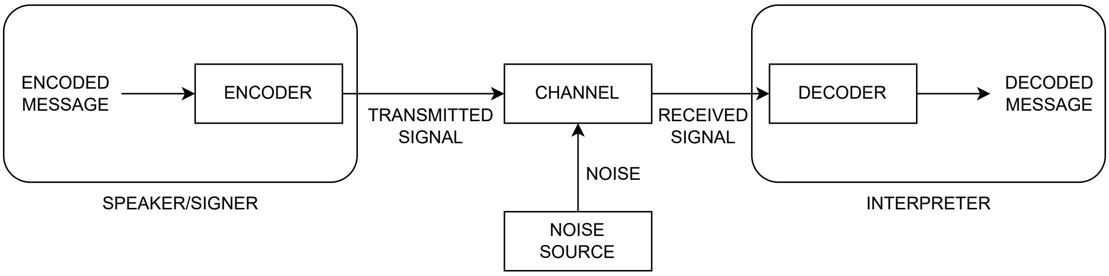


Figure 1.2: Linguistic communication as the encoding and decoding of messages.

Definition 1 (Code)
A code is a bidirectional mapping between signals and messages that is shared by all interacting parties.

Communication thus proceeds by duplicating a message (or "thought") from the mind of a speaker into the mind of an interpreter. Messages are fundamentally *private*, however, since they are speaker-internal and so not straightforwardly accessible to others. Only signals are *public*. Moreover, since the code is purportedly shared by all communicating parties, *noise* becomes the only possible source of misunderstanding or miscommunication. In the absence of noise, the message decoded by the interpreter should be identical to the one encoded by the speaker. In other words: communication success implies *mutual understanding*, characterized here as the identicity of (encoded and decoded) messages.

---

[3] Throughout this thesis, I will employ the term "speaker" to remain consistent with past literature—and notably with Grice (1989) and his notion of "speaker meaning"—but note that most of my considerations apply equally well to sign language and so to both *speakers* and *signers* alike.

1.2.2 Ambiguity against the communicative function of language

Over the past few years, the debate surrounding the communicative function of language has crystallized notably around the question of *ambiguity*. On one side, we find the idea that ambiguity is part of the numerous features that make language a rather poor instrument of communication:

> It is well known that language is in many ways "poorly designed" for communicative efficiency: apart from such ubiquitous phenomena as ambiguity, garden paths, and so on, one core property of languages—recursive embedding with nested dependencies—leads to exponential memory growth and therefore has to be avoided in language use [. . .]
>
> Chomsky (2008, p.136)

The argument seems plain indeed, for:

> If you want to make sure that we never misunderstand one another, for that purpose language is not well designed, because you have such properties as ambiguity.
>
> Chomsky (2002, p.107)

On the other side, we find the idea that ambiguity is not necessarily detrimental to linguistic communication, and that it can even prove *beneficial* to it (e.g. Piantadosi et al., 2012; Santana, 2014; O'Connor, 2015; Mühlenbernd, 2021). Piantadosi et al. (2012) for instance, specifically target the argument from Chomsky (2002, 2008) above in telling us that:

> We argue, contrary to the Chomskyan view, that ambiguity is in fact a *desirable* property of communication systems, precisely because it allows for a communication system which is "short and simple."
>
> Piantadosi et al. (2012, p.281)

Piantadosi et al.'s (2012) argument remains fundamentally rooted in *functionalism*, a view which Bates et al. (1991) define as follows:

> *Functionalism* can be defined as the belief that, the forms of natural languages are created, governed, constrained, acquired and used in the service of communicative functions. So defined, functionalism is the natural alternative to theories of language that postulate a strict separation between structure and function, and/or theories that attempt to describe and explain structural facts *sui generis*, without reference to the constraints on form that are imposed by the goals of communication and the capabilities and limitations of human information processing.
>
> Bates et al. (1991, p.134)

But it actually extends further: it claims not only that language is structured *by* and *for* communication, but also, and more importantly, that it is structured for *efficient* communication (see Gibson et al., 2019; Levshina, 2022, for a recent overview of the argument). The notion of efficiency here mobilized involves the minimization of a "cost-to-benefit ratio" (Levshina, 2022, p.3). The idea is that communication between interlocutors can be considered *efficient* insofar as it proves *effective*—that is, *successful*—all the while requiring minimal processing effort from all communicating parties (Gibson et al., 2019, p.391).

Ambiguity can then be interpreted within this functionalist framework as arising from speakers and interpreters' competing interests—interests which prove to be fundamentally at odds during communication, as Zipf (1949) originally argued. From the perspective of speakers indeed, costless communication implies a maximally ambiguous language, where every possible message can be expressed with just one signal—the same for every message. From the perspective of interpreters, however, costless communication implies a minimally ambiguous language, where each signal maps to a different message.[4] Thus, according to Zipf, if both speakers and interpreters abide to a "principle of least effort" commanding them to minimize their respective processing efforts during communication, they should naturally make language converge to an intermediate level of ambiguity.[5] Which is precisely why, Zipf tells us:

> [W]e may expect that at least some words must have multiple meanings.
>
> Zipf (1949, p.27)

Yet, the thing is that we do not just go around using language in communication by uttering words in isolation. So if it is fair to say that a verb such as *run* in English is ambiguous "in absolute terms", that is, *out of context*—between, e.g., "run a company" and "run a marathon"—the fact of the matter is that in natural conversation, we are much more likely to communicate utterances such as "Camille now runs a company specialized in woodcraft" or "Camille will run the Paris marathon next month" rather than "run" plain and simple.[6] Two utterances which, as you will note, provide important contextual information that makes it possible to disambiguate the meaning of *run* in each case.

Now of course, there is more to ambiguity than just *lexical* ambiguity, but the general point still stands: should we really consider linguistic ambiguity to pose a problem for communication if communicated signals are never, in fact, ambiguous in the first place? Can we really separate the *parts* from the *whole* and argue that linguistic signals are at least partly ambiguous even when communicated signals are never actually processed in part but always in full, with all the surrounding communicative context? In other words: does the presence of linguistic ambiguity alone suffice to make language an ambiguous communication *system*?

---

[4] A version of this argument is present in (Piantadosi et al., 2012, p.281) which relies on (Zipf, 1949, notably chapter 2 and pp.20–22).

[5] See also (Trott & Bergen, 2022) for a more recent version of this argument.

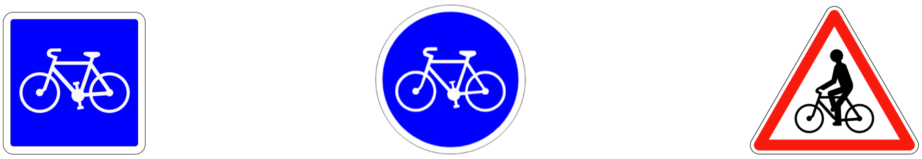
[6] Outside of the imperative "run!" form of course, which is unambiguous anyway.

1.3 Ambiguity against communication: an artificial problem?

1.3.1 Ambiguity against ambiguous systems

Let me take a concrete example to illustrate the distinction I wish to make between *ambiguity* and *ambiguous system*. Consider the system of road signs or traffic signs. The French version—to pick here the variant I am probably most familiar with—is in a certain sense compositional: the *shape* of the road sign appears to make a near-systematic contribution to the overall semantics of the sign (see Figure 1.3).[7] Indeed:

- the *rectangular* shape of the sign signals an *indication* or a *recommendation*: in the case of Figure 1.3a, it advises you to ride your bike on the dedicated cycling road—though it does not compel you to do so;
- the *circular* shape of the sign signals an *obligation*: in the case of Figure 1.3b, it commands you to ride your bike on the dedicated cycling road; and
- the *triangular* shape of the sign signals a *warning*: in the case of Figure 1.3c, it recommends you to watch out for potential surrounding cyclists.



(a) Recommendation (b) Obligation (c) Warning

Figure 1.3: Semantics of the *shape* feature in French road signs.

Yet, that shape feature is also somewhat ambiguous, for it can be found in road signs signaling an *obligation* (Figure 1.4a) or an *interdiction* (Figure 1.4b). In Figure 1.4a indeed, the road sign commands you to ride your bike on the dedicated cycling road while in Figure 1.4b, it specifically forbids to do it on the road where the sign is positioned.

(a) Obligation (b) Interdiction

Figure 1.4: Ambiguity of the *shape* feature in French road signs.

---

[7] See `https://www.siig.fr/fr/docs-diverses/catalogue_signaux_routiers.pdf` for the complete and official list of French road signs with detailed explanations (in French). Whether or not road signs actually form a semantically compositional system does not really matter for the argument I wish to make here, but note that the point is discussed briefly—and argued against—by Szabó (2022, §1.1 and footnotes 2 and 3 specifically) in a USA context.

Still, would we say that the French road sign system is an ambiguous *system*? Probably not, at least insofar as it is precisely supposed to be unambiguous *by design*. Its whole purpose indeed is to communicate clear and unambiguous instructions that can be straightforwardly decoded by every licensed driver. And indeed, both signs in Figure 1.4 are actually unambiguous: they can be both straightforwardly disambiguated if you consider the *set* of features that compose them. In Figure 1.4a, it is the combination of the circular shape of the sign *with* the blue background feature that unambiguously signals an *obligation*; while in Figure 1.4b, it is the combination of the circular shape of the sign *with* the read-and-white background feature that unambiguously signals an *interdiction*.[8]

So ambiguity, from the point of view of the French road sign system at least, is somewhat of a false problem: its presence in a given road sign never actually prevents the proper decoding of the corresponding instruction. But if it could be qualified as a "false problem" here, it is also because it arises exclusively from arbitrarily considering road sign features in isolation, and thus from somewhat artificially isolating *parts* of the system from the *whole*. Yet, only the "fully-formed" and otherwise unambiguous signs actually correspond to the *real* use cases of the system, and only those can therefore be used to arbitrate on the true nature of French road signs as an ambiguous *system*. In the end, the presence of ambiguity alone amongst the compositional features of French road signs does not suffice to make French road signs an ambiguous system as a whole.

For the question of language and communication then, such considerations imply two things. First, that the sole presence of ambiguous linguistic forms does not suffice to make language an ambiguous *system*. Second, that the ambiguous nature of language as a communication system can only be settled by considered (fully-formed) *communicated signals*. Under the framing of the code model, I thus propose to define the notion of "ambiguous communication system" as follows (Definition 2).

> **Definition 2 (Ambiguous communication system)**
> A communication system is ambiguous if and only if there exists a one-to-many mapping between a communicated signal and its corresponding messages.

1.3.2 Putting linguistic ambiguity back into communicative context

In his book on *Language and Communication*, Miller (1951) introduces the problem of ambiguity as follows:

> Consider the verb 'take'. In the Oxford English Dictionary there are 317 definitions of this word, 317 verbal alternatives that can be substituted for it. This is a confusing state of affairs. [...] Why do people tolerate such ambiguity? The answer is that they do not. There is nothing ambiguous about 'take' as

---

[8] Note that this disambiguation is not the sole product of the *red-and-white background* feature since, in Figure 1.3c, we find it to be signaling a *warning* when found in combination with a triangular shape, and not an *interdiction*. So, once again, it is the *combination* of features that disambiguates the parts and makes the whole sign unambiguous. Here, the *red-and-white background* feature could equally be said to be "ambiguous".

> it is used in everyday speech. The ambiguity appears only when we, quite arbitrarily, call isolated words the unit of meaning.
>
> Miller (1951, pp.111–112)

His argument proves crucial to us here, for it actually suggests that linguistic ambiguity may arise solely from our arbitrarily divorcing communicated signals from their communicative context for the needs of linguistic inquiry.[9] In this case, the argument seems plain, since the meaning of *take* as "accept" in (1.1 a) or as "cost" in (1.1 b) can clearly be distinguished thanks to their surrounding (linguistic) context.

(1.1) (a) I think you should take the offer.
(b) How much will it take to buy this house?

The problem with the code model, however, is that it leaves us little room to accommodate this notion of "context" outside of communicated signals—given that the code must be shared by all communicating parties. This means that communicated signals must necessarily correspond to *(linguistic) signal–context pairs* rather than *bare* linguistic signals. The problem with this notion of *context*, then, is that we do not know exactly where it stops, or what it is made of. First of all—and contrary to the example of road signs in the previous Section 1.3.1—we do not really know what constitutes a "fully-formed" communicated signal or what sets the boundaries of those signals in the case of linguistic communication. In the corresponding entry of his *Dictionary of Linguistics and Phonetics*, Crystal (2008) makes clear that what we usually take to be the standard unit of linguistic communication—the *utterance*—is always somewhat vague and arbitrary (see Definition 3).

> **Definition 3 (Utterance)**
> A term used in linguistics and phonetics to refer to a stretch of speech about which no assumptions have been made in terms of linguistic theory (as opposed to the notion of sentence, which receives its definition from a theory of grammar). In principle, it is a physically definable, behavioural unit, capable of definition in everyday terms. One commonly used definition refers to a 'stretch of speech preceded and followed by silence or a change of speaker'. But it has proved very difficult to construct a satisfactory definition. The definition just given, for instance, applies equally to a one-word response and a sermon, and attempts have been made to produce a more restricted definition, using such features as pause, rhythm, breath patterns, pitch movement, etc. The analogous term in the study of writing is text.

---

[9] If so, then linguistic ambiguity could probably be considered a product of what the physical science call the "observer effect", since it would be a direct product of the methods used to "observe" (in this case, analyze or study) language itself (see Sassoli de Bianchi, 2013, for an overview of the observer effect in physics). The point being that in linguistics, we always need to break down signals into smaller entities so as to study them in isolation, but that this methodological necessity actually creates "interferences" with our object of study and artificially generates the phenomenon of ambiguity in turn. In other words, ambiguity would just be a byproduct of our study of language.

Adapted from Crystal (2008, pp.505–506)

Second, the "context" involved in linguistic communication often extends much beyond *linguistic* context proper—unlike in example (1.1)—and may cover a wide range of *non*-linguistic modalities as well.

Both of those considerations have a direct impact on how we approach the problem of ambiguity in communication. I could, for instance, argue that the example in (1.2 b) constitutes a genuine case of communicative ambiguity: given that it is here formalized as a single utterance, it should be treated as a "fully-formed" communicated signal containing a persistent referential ambiguity in the demonstrative *that*.

(1.2) (a) Camille did not come to work yesterday.
(b) That's strange.

But that would be completely artifactual of arbitrarily segmenting the utterance in a way that would exclude from the communicated signal everything that has been said *before* in conversation, such as (1.2 a). Treat the same example as a single utterance now, like in (1.3), and the communicative ambiguity vanishes, given that the reference of the demonstrative "that" is clearly disambiguated by the communicated signal as a whole.[10]

(1.3) Camille did not come to work yesterday. That's strange.

In practice, linguistic communication always takes place under a particular *situational* context where what has been said before in conversation—as well as the time or place under which communication takes places, the beliefs, knowledge or presuppositions of interlocutors, and their awareness of their surrounding environment—may all play a key role in the interpretation of communicated messages. In (1.3), for instance, it is the interpreters' knowledge of the situational context that tells them what date "yesterday" actually refers to. In that sense, the situational context is very much *non*-linguistic as well: if I ask you (1.4) while pointing at a box in your hand, it is my gestures that tell you exactly what I mean by "that".

(1.4) What's that?

Thus, in all those cases, arbitrarily separating the communicated signal from its multimodal situational context creates cases of ambiguity which are completely artificial from the point of view of communication.

The specific modality under which we study language also seems to imply "tampering" with communicated signals to some extent. Think of how, for instance, the mere fact of

---

[10] How long would have to be the pause denoted by the comma in natural conversation for that segmentation/unification of the utterance to be considered valid? Precisely, the formal definition of "utterance" in Definition 3 does not say. Of course, one could also perfectly consider that (1.3) always characterizes a single piece of discourse composed of two separate utterances, but the question would then be what constitutes the boundaries of "discourse" (see Taylor & Cameron, 1987; Schiffrin, 1994; Stede, 2012, for an overview of the question).

writing down those linguistic examples artificially removes prosody. Yet, there are many situations in verbal communication where we can actually resort to prosody to disambiguate linguistic signals. As speakers we can, for instance, modulate the amplitude, pitch contour or duration pattern of an utterance to orally distinguish "$2 + (3 \times 4)$" from "$(2 + 3) \times 4$" (Streeter, 1978). We can also use prosody to overcome cases of *structural* or *syntactic* ambiguity as in (1.5), for instance, where the prepositional phrase "with the telescope" is ambiguous as to whether it is attached to the verb phrase "saw the man", as in Figure 1.5, or to the noun phrase "the man", as in Figure 1.6.

(1.5) I saw the man with the telescope.

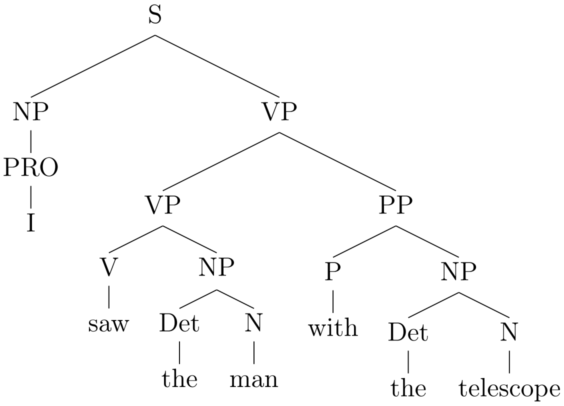


Figure 1.5: I saw the man [with the telescope]

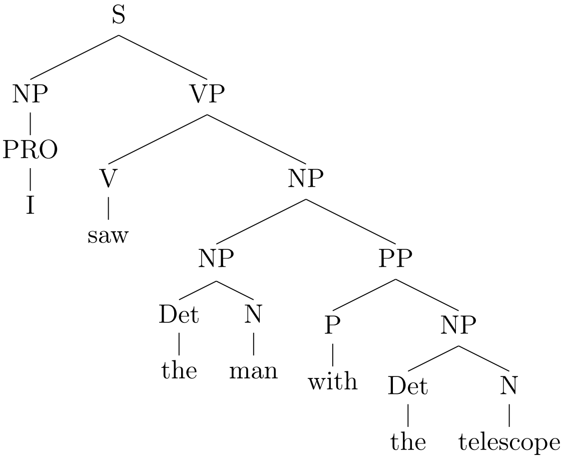


Figure 1.6: I saw [the man with the telescope]

Yet as speakers once again, we can actually clear up that ambiguity by marking a prosodic boundary after "man", as in (1.6 a), or after "saw", as in (1.6 b), by lengthening the noun or the verb and the pause immediately following it (see Kraljic & Brennan, 2005, for details).

(1.6) (a) I saw the man . . . with the telescope.

(b) I saw . . . the man with the telescope.

All the observations developed so far rejoin the more general argument already made by others, that ambiguity need not be seen as a problem for communication if context actually enables the interpreter to retrieve the encoded message (e.g. Piantadosi et al., 2012; Santana, 2014).

However, it is one thing to argue that context *can* be informative about the intended message, and yet another to argue that it is *always* informative enough in practice. Similarly, it is one thing to argue that ambiguity may not constitute a definite obstacle to successful communication, and yet another to argue that it never hinders the communication process in any way. Linguistic ambiguity would still make language an inefficient communication system if it systematically required extra processing effort from all communicating parties—even if communication were to prove systematically successful nonetheless.

Yet, even that is not so clear: moving past the intuition—partly supported by empirical evidence—that linguistic ambiguity should necessarily be costly for interpreters, the psycholinguistics literature on the topic depicts a much subtler reality (see Rodd, 2018; Grant et al., 2020, for overviews in the context of lexical ambiguity and structural/referential ambiguity respectively). Traxler et al. (1998), for instance, show that the ambiguous sentence in (1.7 a) actually proves easier to process than its unambiguous counterpart in (1.7 b).

(1.7) (a) The son of the driver that had the mustache was pretty cool.
(b) The car of the driver that had the mustache was pretty cool.

Much of the work reviewed by Rodd (2018) and Grant et al. (2020) point in a similar direction, namely, that having to entertain a multiplicity of options does not necessarily imply extra processing effort for the interpreter. But their work also stress a much more fundamental consideration: that our very intuition about the processing cost of linguistic ambiguity may be biased by years of psycholinguistics research investigating language in experimental setups divorced from communicative context.

Indeed, an important argument that is often mobilized by those in favor of not seeing ambiguity as posing a definite problem for communication is that interlocutors do not seem to care or pay much attention to it—let alone try to avoid it (see Wasow, 2015, and references therein). Chomsky (1965) himself argues that our ability to interpret utterances in context is so straightforward that we can even fail to detect syntactic ambiguity when it occurs:

> If a sentence such as "flying planes can be dangerous" is presented in an appropriately constructed context, the listener will interpret it immediately in a unique way, and will fail to detect the ambiguity. In fact, he may reject the second interpretation, when this is pointed out to him, as forced or unnatural (independently of which interpretation he originally selected under contextual pressure).
>
> Chomsky (1965, p.21)

In psycholinguistics experiments, however, ambiguity avoidance appears to depend much on whether or not subjects are tested in clear communicative contexts. Allbritton et al. (1996), for instance, show that speakers do not reliably generate prosodic cues to help interpreters disambiguate otherwise structurally ambiguous sentences—such as in (1.6)—but they only test subjects in a monologue context and through read-aloud tasks. When Schafer et al. (2000) investigate the very same question but through a cooperative game this time—which requires speakers and interpreters to interact with clear communicative goals—they find that speakers *do* produce those prosodic cues and that they do so *systematically*, even when the structural ambiguity could have been resolved by other means.

Regarding now the (dis)advantage of ambiguity observed in traditional psycholinguistics experiments, Swets et al. (2008) bring us interesting nuance. They show that cases of structural ambiguity such as (1.7) can be turned into an advantage/disadvantage alternatively, depending on how specifically the structural ambiguity is targeted by the experimental setup. Concretely, they show that if you ask subjects a specific question such as "Did the son of the driver have a moustache?", you find (1.7 a) to be harder to process than (1.7 b). But if you ask subjects a general question such as "Did anyone have a moustache?", you find (1.7 a) to be easier to process than (1.7 b). In the end, the presence of structural ambiguity alone does not appear to be an advantage/disadvantage *in itself*. To definitely impact processing cost in communication, it needs to be directly relevant to the communicative situation at hand.

Finally, the study of Swets et al. (2008) suggests something fundamental, namely, that what we take to be "ambiguous linguistic forms" is very much conditioned on the theoretical choices that we make in the first place. Those theoretical choices condition our empirical expectations, in turn—such as the fact that processed linguistic signals should always correspond to complete and well-defined syntactic parses (see Chater et al., 1998), which is precisely the assumption that underlies our appreciation of structural or syntactic ambiguity. But maybe communicated messages need not always match complete and optimal syntactic parses, maybe they need not be "truth-evaluable" either, have definite conceptual extensions, or correspond to neat categories. Maybe the very existence of linguistic ambiguity is nothing but an artifact of linguistic inquiry altogether, of the theoretical devices that we mobilize to study language, of our methodological and experimental choices, and of a practice of linguistics in general which often requires that we split fully-formed linguistic signals into parts and divorce them from their communicative context so as to study them more easily.[11] If so then ambiguity may be but an artificial problem for communication after all.

1.3.3 Making sense is a process

Is language an ambiguous communication *system*, then? If interpreted in a "strict" sense, Definition 2 implies that interpreters should somehow "get stuck" in decoding when confronted with an ambiguous communicated signal, as they should not be able to single out any particular messages from the multiplicity of options entailed by the one-to-many mapping. But what would it actually *mean* in practice for interpreters to "get stuck" in decoding because of an ambiguous communicated signal?

Stepping aside from the problem of ambiguity for a moment, we can observe that there does seem to exist configurations in language that get us "stuck" in parsing—in the sense that they can prove nearly impossible for us to process. Think of the example of center-embedded sentences, for instance, such as (1.8) taken from (Bever, 1970/2013, p.58), where our processing ability degrades progressively each time an embedding is added, until the sentence becomes nearly impossible to parse (see also Miller & Isard, 1964, for details).

---

[11] A practice itself which often takes place within an arbitrary separation between *structure* and *use* which functionalism (not to mention *usage-based* approaches to linguistics) precisely criticizes (see Langacker, 1987, for the latter).

(1.8) The dog the cat the fox was chasing was scratching was yelping.

Could we, in a similar fashion, find or construct examples of strictly ambiguous signals that we would find nearly impossible to process, in the sense of that we would find it nearly impossible to recover any kind of message at all?

Imagine a situation where we would both walk past a bank (the financial building) along a river bank. Without any further (situational) context, I would tell you:

(1.9) Can you see the bank?

How exactly would you make sense of (1.9)? Would you actually "get stuck" in interpretation and prove unable to decode any message at all? I would argue not. We could imagine that you would reply "which one?" and your answer would only stress the limitations of a code model that characterizes communication as a passive sender–receiver game rather than an interactive process.[12] We could also imagine that you would interpret the ambiguity of my utterance somewhat "logically" and reply "yes" given that you would decode it as *Can you see the financial building and/or the river bank?*. This time, your answer would probably emphasize that communicated messages need not always be "precise" (see Clark, 1997, and his "dogma of determinate meaning") and that communication is often not a yes–no matter but a matter of *degree* (Wilson, 2018, p.189). But you could also infer that I was just trying to make a joke and laugh. In any case, the point is that there seems to be a qualitative difference between (1.8) and (1.9) in that the former seems to get us "stuck" in processing in a way that the latter does not.

In fact, cases of performance limitation in language processing are all the more interesting to us here than they illustrate how such configurations where we actually get "stuck" in interpretation remain fundamentally rare. If you consider the seminal example of *garden path* from (Bever, 1970/2013, p.40) in (1.10), it also shows how the human mind is always structured so as to "do its best using whatever it has" (Wang, 2019, p.16).

(1.10) The horse raced past the barn fell.

In parsing (1.10) indeed, we do not just stop and give up upon encountering the word "fell"—even though it invalidates our previous interpretation that *The horse raced past the barn*. We try again and *re*-parse the sequence until we get *some* form of representation that is compatible with the full linguistic signal (Frazier & Fodor, 1978; Frazier, 1979).

We see here how those considerations sketch what could be qualified as some sort of "drive to make sense" from the human mind—not only of linguistic signals but also of the surrounding world as a whole and of external stimuli in general. To illustrate that intuition, imagine a situation where I would come to you and say:

(1.11) Bazinga.

---

[12] An interactive process probably best formalized as the *alignment of information states* rather than *information transfer* actually (e.g. Pickering & Garrod, 2006, 2013, 2021).

The fact that (1.11) may not constitute a semantically "valid" or "well-formed" English expression would probably not stop you from trying to make sense of it and ask: what could I have possibly meant here? What was the message I was trying to convey? What was my communicative intention? Did I mean to tell a joke? To express an emotion? Was I trying to scare you? Or just speaking a foreign language? Chances are you will assume that I must have meant *something* by (1.11)—or that, at the very least, I must have produced those sounds for a *reason*. So even if you do not decode a precise message—let alone *my* message—I would very much argue that you will strive to make sense of this situation, even if it implies concluding that I did not mean anything at all and that I was just being "crazy" (appeal to "madness" probably being the convenient solution we too often mobilize to account for human behavior we cannot otherwise explain).

Maybe there is more to communication than language, and maybe communication is much more inferential than the code model suggests. Be that as it may, what the field of pragmatics teaches us first is that *context* plays a central role in communication: to make sense *is* to contextualize—to construct a relevant background of information making a communicated signal interpretable in practice, that is, "mappable" to a particular message. And as poetry and literature often suggest, the only limitation to that ability to contextualize signals and make sense of language in general appears to be the limits of our own imagination. The best example I can think of to illustrate that point is the *colorless green ideas* of (Chomsky, 1957/2002, p.15), which probably constitutes the most famous example of modern linguistics. Chomsky originally uses this example to build a case for a notion of "grammaticality" distinct from that of "meaningfulness"—his point being that while both examples in (1.12 a) and (1.12 b) are equally "nonsensical", only the former is actually *grammatical*.

(1.12) (a) Colorless green ideas sleep furiously.
(b) Furiously sleep ideas green colorless.

Yet, the meaninglessness of (1.12 a) may have been glossed over too quickly, for in the 1990s indeed, Stanford University organized a literary competition with the explicit instruction to devise a context that would make that *colorless* example in (1.12 a) meaningful.[13] The instructions went as follows:

> For the Literary competition set on Christmas Eve you were asked to compose not more than 100 words of prose, or 14 lines of verse, in which a sentence described as grammatically acceptable but without meaning did, in the event, become meaningful. The sentence, devised by Noam Chomsky, was: colourless green ideas sleep furiously.

And the winning poem was (colorless example in bold):

> Thus Adam's Eden-plot in far-off time:
> Color-rampant fowers, trees a myriad green;

---

[13] See `https://www.linguistlist.org/issues/2/2-457/#2` for details.

> Helped by God-bless'd wind and temp'rate clime.
> The path to primate knowledge unforeseen,
> He sleeps in peace at eve with Eve.
> One apple later, he looks curiously
> At the gardens of dichromates, in whom
> **colourless green ideas sleep furiously**
> then rage for birth each morning, until doom
> Brings rainbows they at last perceive.
>
> D. A. H. Byatt

This literary experience proves particularly interesting to us here as it provides a concrete illustration of a case where this contextualization process gets "materialized" linguistically. But if the poet helps us make sense of that particular utterance by surrounding it with a particular linguistic context, ultimately it is *us*, the interpreters, who make sense of that utterance in context. For it is the interpreters who make sense of language and not language that carries meaning on its own:

> Expressions do not mean; they are prompts for us to construct meanings by working with processes we already know. In no sense is the meaning of [. . .] any utterance "right there in the words." When we understand an utterance, we in no sense are understanding "just what the words say"; the words themselves say nothing independent of the richly detailed knowledge and powerful cognitive processes we bring to bear.
>
> Turner (1991, p.206)

In sum, *making sense is a process*, where context is not something that is "given" but rather something that is "built" and so not a *property* of communicated signals but rather an *ability* of interpreters—and an ability that appears to be practically unbounded.

If our ability to make sense of communicated signals is truly unbounded, then we should never actually find ourselves "stuck" in decoding and should thus be able to safely conclude that language is not an ambiguous communication *system*, insofar as it should always prove possible to decode *at least one* message from a communicated signal—ambiguous or not. However, that would not suffice to conclude that language is an *efficient* communication system for as we saw in Section 1.2.2, *efficiency* presupposes *effectiveness* and communicative effectiveness itself requires that interpreters decode *the* message encoded by the speaker and not just *any* message. So the question we must now turn to is this: is language actually an *effective* communication system?

1.4 Is language an effective communication system?

1.4.1 Underdetermination as ambiguity applied to communication

> Of course I can use the words ‘this man’ to designate now this man, now that man. But still on each single occasion I mean them to designate just one man. The sentences of our everyday language leave a good deal to guesswork. It is the surrounding circumstances that enable us to make the right guess. The sentence I utter does not always contain everything that is necessary; a great deal has to be supplied by the context, by the gestures I make and the direction of my eyes.
>
> Frege (1914/1979, p.213)

Frege’s quote above nicely summarizes some of the important points I have been trying to develop throughout Section 1.3. First, linguistic signals are not necessarily *autonomous*: they sometimes need to be paired with *situational* and potentially *non-linguistic* context in order to be interpreted. Second, linguistic ambiguity—pervasive as it may be—is never *strict* communicative ambiguity: we do seem to be able to get by linguistic ambiguity in communication and to always interpret at least *a* message, if not *the* message intended by the speaker. Third, we do not manage to get by linguistic ambiguity in any random way: we *construct* an adequate context to be paired with the linguistic signal and do so by appealing to our *inferential* abilities—what Frege (1914/1979) calls “guesswork” here.[14]

However, I have acted so far as if all those considerations could be easily accommodated within the code model but this is actually far from given. Indeed, the code model is not really conceived to function with communicated signals that are not autonomous in the first place, and accommodating the very notion of context within it remains quite a conceptual challenge (as we briefly saw in Section 1.3.2). Bypassing the problem by “refactoring” communicated signals as (linguistic) signal–context pairs rather than bare linguistic signals is, in a certain sense, nothing but a convenient “hack” in the theory (to employ here a technological metaphor in line with the original spirit of the code model) and as we will see in Section 1.5.1, it is most likely a misguided one. But the biggest challenge to the code model actually comes from this necessity to appeal to *inferences* in communication, for it actually calls into question the very *coded* nature of human communication, and so the very relevance of the code model itself.[15]

---

[14] Though note here that some may argue that inferences are actually distinct from “guesswork” in the sense that they precisely do not rely on “betting” on a possible meaning or interpretation but on deriving it logically and/or rationally. But I will leave this particular nuance aside for my present purpose.

[15] In *The Oxford Dictionary of Pragmatics*, Huang (2012) defines the notion of “inference” as follows:

> As one of the basic forms of reasoning, inference is a process of accepting a statement or proposition (called **conclusion**) on the basis of the (possibly provisional) acceptance of one or more other statements or propositions (called the **premises**). (Huang, 2012, p.156)

He also defines “reasoning” as follows:

> A systematic mental process in which implications among propositions are evaluated, and a conclusion is drawn from a set of premises. One of the basic forms of reasoning is infer-

The problem with the code model is that there appears to be cases of human communication that are actually *purely* inferential and as such not coded *at all.* If I ask you (1.13) and you reply to me by waving a box of aspirin, for instance, I will straightforwardly interpret that you mean to tell me that you are not feeling well (see Sperber & Wilson, 1986/1995, pp.25–26, for the original example).

(1.13) How are you feeling today?

Yet, as Sperber & Wilson (1986/1995) stress, this behavior is not coded: there is no rule or convention that tells us here that waving box of aspirin should systematically be interpreted as meaning that one is not feeling well. In a different context, the same gesture could very well mean “I have some aspirin if you need”, or “I'm ready for our trip!”.

Maybe human communication in general is much more inferential than we think—there is probably more to communication than language anyway—but the definite problem with the code model is that linguistic communication *itself* also proves to be much more inferential than we think—as the Gricean turn in pragmatics fundamentally emphasized (see Grice, 1989, for an overview). In his seminal *Meaning* paper, Grice (1957) invites us indeed to treat linguistic communication as a problem of *intention resolution*, where to *mean* something by employing a particular expression is to intend to produce an effect in an audience by having that audience recognize that intention (see Definition 4).

> **Definition 4 (Communication as intention resolution)**
> “A meant something by x” is (roughly) equivalent to “A intended the utterance of x to produce some effect in an audience by means of the recognition of this intention”.
>
> Grice (1957, p.384)

Crucial to Grice's proposal, then, is the fundamental distinction between *speaker meaning* and *sentence meaning*—originally detailed in (Grice, 1968)—and the idea that there is more to communication than what is said. If you ask me (1.14 a) and I reply (1.14 b), for instance, what I imply is that “No, I do not want coffee”, so that what I *mean* (i.e. the “speaker meaning”) is actually different from what I *say* (i.e. the “sentence meaning”).

(1.14) (a) Do you want coffee?
(b) Coffee gives me heartburn.

Under Grice's inferential account of communication, messages are therefore *intentions* which correspond to a particular type of meaning—the speaker meaning—and the critical point of Grice's distinction is that sentence meaning fundamentally underdetermines speaker meaning—which is why, Grice (1975) tells us, speakers should abide to a “cooperative principle” (Definition 5) as well as a set of “conversational maxims” (Definition 6) to make sure that they are correctly understood by their audience.

---

ence. (Huang, 2012, p.262)

Definition 5 (Cooperative principle)
Make your contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.

Grice (1975, p.45)

Definition 6 (Maxims of conversation)

1. Quantity maxims:
   a) Make your contribution as informative as is required (for the current purposes of the exchange).
   b) Do not make your contribution more informative than is required.
2. Quality maxims:
   a) Do not say what you believe to be false.
   b) Do not say that for which you lack adequate evidence.

   Supermaxim: Try to make your contribution one that is true.
3. Maxim of Relation: Be relevant.
4. Manner maxims:
   a) Avoid obscurity of expression.
   b) Avoid ambiguity.
   c) Be brief (avoid unnecessary prolixity).
   d) Be orderly.

   Supermaxim: Be perspicuous.

Adapted from Grice (1975, pp.45–46)

How inferential is linguistic communication exactly? This is where different approaches to pragmatics diverge. To better understand their differences, it is important first to introduce what Recanati (2004, §1.1) calls the "basic triad"—which further decomposes the concepts of sentence meaning and speaker meaning into *sentence meaning*, *what is said* and *what is implicated*. The distinction goes as follows: if you ask me (1.15 a) and I reply (1.15 b), for instance, what I really mean in that context (i.e. what is implicated/what is meant/the speaker meaning) is that *yes, I have already tried cheese fondue*—which you will straightforwardly infer if you know that cheese fondue is a typical Swiss dish.

(1.15) (a) Have you ever tried cheese fondue?
(b) I live in Switzerland.

But the sentence in (1.15 b) also has, arguably, a “linguistic meaning” (i.e. sentence meaning) that is independent of the context of utterance, and that could be roughly characterized as (1.16 a). Meanwhile, what is (literally) said by (1.15 b)—which can be roughly characterized as (1.16 b)—depends on what the indexical “I” refers to in that context.

(1.16) I live in Switzerland.

(a) *Sentence meaning*: The speaker lives in Switzerland.

(b) *What is said*: Alexandre Kabbach lives in Switzerland.

(c) *What is implicated*: Yes, Alexandre Kabbach has already tried cheese fondue.

Different approaches to pragmatics then regroup those notions differently, notably depending on how inferential they take linguistic communication to be: *literalists* (and *minimalists* in particular) take what is said to be part of the literal meaning of the proposition expressed—alongside sentence meaning—and to remain pretty close to sentence meaning (hence the term “minimalist/minimalism” which refers to the assumption that the distance between sentence meaning and what is said should be minimal). *Contextualists*, on the other hand, take what is said to belong to speaker meaning alongside what is implicated (see Recanati, 2004, p.6).

Now, in the example (1.15) above there is a clear gap between what is said and what is implicated by (1.15 b) (and thus between the literal meaning of the utterance and the speaker meaning). But we could very well imagine cases where that gap would be pretty small, if not actually null, in that what is meant (or the speaker meaning) would be identical to what is (literally) said: for instance, if you ask me (1.17 a) and I reply (1.17 b).

(1.17) (a) Where do you live?

(b) I live in Switzerland.

Yet, even in (1.17 b) one would have to derive who “I” and “you” actually refer to, so that if what is said underdetermines speaker meaning, sentence meaning also underdetermines what is said (see Carston, 2002, notably §1, for an overview). What is more, the role played by inferences in determining what is said does not stop at indexical or linguistic ambiguity resolution. Sentence meaning is often much more sketchy and fragmentary than what literalists assume, and often too schematic to constitute a full proposition that would be truth-evaluable—contrary to what literalists expect. In (1.18 a), for instance, one has to decide what the sea is too cold for before they can assign that proposition a definite truth-value; and in (1.18 b), one has to decide whether the book is actually difficult to read, write, or sell, for instance, and by comparison to what (see Sperber & Wilson, 2012, p.9 for the original discussion over those examples).

(1.18) (a) The sea is *too cold*.

(b) That book is *difficult*.

Pragmatic inferences thus appear to be necessary not only to derive what is implicated by a communicated signal, but also what is said.[16]

Anyhow, if pragmatic inferences are so pervasive in linguistic communication, it is first and foremost because *communicated signals structurally underdetermine their messages*. The fact of the matter indeed is that:

> Most—arguably all—sentences of any human language [...] underdetermine their interpretation.
>
> Sperber & Origgi (2010, p.124)

This structural underdetermination of messages means that linguistic communication always leaves open a multiplicity of options available as to which messages should be interpreted, and that it does so by default. That is, underdetermination is not a byproduct of noise in the transmission channel, or of factors that are extrinsic to the communication system itself: it is an inherent property of language and linguistic communication. What is more, this idea of leaving open a "multiplicity of options" between signals and messages brings back in full the very notion of *communicative ambiguity*—formally defined as the one-to-many mapping between signals and messages—so that we could say, in a sense, that *underdetermination* is nothing more than *ambiguity applied to communication*.

Having clarified the terms of the debate, we can now look back at the original argument of Chomsky (2002, 2008) and give it due credit. For indeed, there does seem to be a problem with ambiguity in communication, just not with the traditional notion of "linguistic" ambiguity. The problem indeed does not lie with those "ambiguous forms" which are easily resolved in communication by appealing to situational context, and probably constitute an artificial problem from the point of view of communication anyway. No, the real problem lies with *underdetermination*, which constitutes the first major obstacle on the road to communication success and communicative effectiveness: how can we make sure indeed that the interpreted message always corresponds to the intended one?

1.4.2 Subjectivity against effective communication

Structural underdetermination means structural communicative ambiguity, and yet as we saw in Section 1.3.3 language is not an ambiguous communication is the strict sense. Therefore, underdetermination should not constitute a definite obstacle to the interpretation of *at least one* message. But could it constitute a definite obstacle to communication success nonetheless? How can we make sure indeed that the interpreted message is in fact *the* message the speaker intended to convey given underdetermination? How can we guarantee *mutual understanding* and the *identicity of messages* that goes with it in that context?

Grice (1975) offers us a possible solution to that problem with his cooperative principle and his maxims of conversation. The problem with Grice's approach, however, is that it

---

[16] Note that contextualists ultimately argue for dropping the concept of "what is said" altogether, given how difficult it is to actually separate "what is said" from "what is implicated" (see Sperber & Wilson, 2012, §1.2, for a full discussion).

is fundamentally *normative*: it tells us what speakers *ought to* do to be understood in conversation, and not what they *actually do* in practice.[17] As a matter of fact, speakers do seem to violate the maxims on certain occasions, especially in the rhetorical use of language (though not only in such cases, see Wilson & Sperber, 2002, for a general overview of the argument). *Irony*, for example, violates the maxim of quality which commands us to make our contribution one that is true: uttering (1.19) when it is actually pouring rain outside implies saying something that is *literally false*.

(1.19) What a great weather for a picnic!

Metaphors, on the other hand—such as example (1.20) from (Wilson & Sperber, 2002, p.587)—violate the maxim of quantity which commands us to make our contribution informative: given that *nothing is said*, no information is provided at all.

(1.20) The leaves danced in the breeze.

To remedy the shortcomings of Grice's normative approach, Relevance Theory (Sperber & Wilson, 1986/1995) proposes a *descriptive* approach to human communication centered on the notion of *relevance*. An utterance is deemed all the more relevant than it yields positive cognitive effects to the interpreter, all the while requiring little cognitive effort to be processed (see also Sperber & Wilson, 2012, p.6, for details). According to Relevance Theory, human cognition and communication are centered around two fundamental principles: a "*cognitive* principle of relevance" (Definition 7) and a "*communicative* principle of relevance" (Definition 8).

> **Definition 7 (Cognitive principle of relevance)**
> Human cognition tends to be geared to the maximization of relevance.

---

[17] I use the term "normative" rather than "rationalist" to characterize Grice's approach as the opposition between Grice and Relevance Theory is usually framed under the normative/descriptive dichotomy (see Sperber & Wilson, 1986/1995; Wilson & Sperber, 2012; Ludlow, 2014; Korta & Perry, 2020, pp.271;6;12;93 respectively, for typical examples). For instance, Sperber & Wilson (1986/1995) tell us that:

> It would be a mistake to read the presumption of optimal relevance, in either the early or the revised version, as describing a goal that rational communicators should achieve. Unlike Grice's maxims, neither the principle nor the presumption of relevance is presented as a goal to be pursued or a rule to be followed by the communicator. **The (Second) Principle of Relevance is a descriptive (as opposed to normative) claim about the content of a given act of ostensive communication**. It claims that part of that content is a presumption that this very act of communication is relevant to the addressee. (Sperber & Wilson, 1986/1995, p.271, my emphasis in bold)

And Ludlow (2014) otherwise frames the opposition quite clearly as follows:

> Relevance Theory thus contrasts with Grice's approach in two ways. First, it subsumes all of Grice's maxims under the "least effort" principle (which they confusingly call "relevance"), and second it represents a shift away from thinking about pragmatics as a process involving rational communicative principles and repositions it as a lower-level process in cognitive psychology. For Grice, the maxims were normative. In Relevance Theory, the principle of relevance is only normative in the sense a biological norm is. **Relevance Theory is, at bottom, a descriptive project and not a normative project.** [Footnote: "Thanks to Dan Sperber and Deirdre Wilson for discussion of this point"] (Ludlow, 2014, p.93, my emphasis in bold)

Sperber & Wilson (1986/1995, p.260)

Definition 8 (Communicative principle of relevance)
Every act of overt communication conveys a presumption of its own optimal relevance.

Sperber & Wilson (1986/1995, p.260)

Two additional specifications come to supplement the theory. The first one is what is called the "presumption of optimal relevance" (Definition 9) which mirrors my "drive to make sense" introduced in Section 1.3.3—although it takes a slightly different perspective on the object of study since the presumption of optimal relevance is conceived as a property of the signal while the drive to make sense is conceived as an ability of the interpreter.

Definition 9 (Presumption of optimal relevance)

a. The utterance is relevant enough to be worth processing.

b. It is the most relevant one compatible with the communicator's abilities and preferences.

Sperber & Wilson (2012, p.7)

The second specification is what is called the set of "relevance-guided comprehension heuristic" (Definition 10) which specifies further the underlying forces driving human cognition and interpretation, and also reminds us of the principle of least effort of Zipf (1949) introduced in Section 1.2.2.[18]

Definition 10 (Relevance-guided comprehension heuristic)

a. Follow a path of least effort in constructing an interpretation of the utterance (and in particular in resolving ambiguities and referential indeterminacies, in going beyond linguistic meaning, in supplying contextual assumptions, computing implicatures, etc.).

b. Stop when your expectations of relevance are satisfied.[19]

Sperber & Wilson (2012, p.7)

Relevance Theory takes as a starting point a Gricean perspective on the inferential nature of human communication rooted in a fundamental critique of the code model and its limitations (see Sperber & Wilson, 1986/1995, notably ch.1 on *Communication*). Contrary to the Gricean approach, however, Relevance Theory considers that inferences play a much more pervasive role in communication and are not restricted to deriving what

---

[18] Note that the presumption of optimal relevance is originally found in (Sperber & Wilson, 1986/1995, p.267), though under a slightly different phrasing, and that the relevance-guided comprehension heuristic is originally found in (Sperber et al., 1995, p.51).

[19] In (Wilson & Sperber, 2004/2006, p.613) they add: "(or abandoned)".

is *implicitly* communicated in conversation but also what is *explicitly* communicated.[20] Under Relevance Theory, pragmatic enrichment is both necessary, and systematic: every communicated signal must be contextualized in order to be interpreted. Contextualization then proceeds by supplementing each communicated signal with a particular context, which Sperber & Wilson (1986/1995) define as follows:

> The set of premises used in interpreting an utterance (apart from the premise that the utterance in question has been produced) constitutes what is generally known as the *context*. A context is a psychological construct, a subset of the hearer's assumptions about the world. It is these assumptions, of course, rather than the actual state of the world, that affect the interpretation of an utterance. A context in this sense is not limited to information about the immediate physical environment or the immediately preceding utterances: expectations about the future, scientific hypotheses or religious beliefs, anecdotal memories, general cultural assumptions, beliefs about the mental state of the speaker, may all play a, role in interpretation.
>
> Sperber & Wilson (1986/1995, pp.15–16)

Relevance Theory thus treats "context" not as something that is *given* in conversation but as something that is *built* by interpreters (Sperber & Wilson, 1986/1995, §3.3)—just as I did in Section 1.3.3—but its characterization of context also fundamentally brings in *subjectivity*. For beliefs, knowledge, memories or assumptions all constitute as many possible attestations of the idiosyncrasy of an interpreter's mental content indeed, and if we have every reason to believe that context—and thus ultimately interpretation—may differ from one interpreter to the next, it is precisely because:

> True, all humans are constrained by their species-specific cognitive abilities in developing their representation of the world, and all members of the same cultural group share a number of experiences, teachings and views. However beyond this common framework, individuals tend to be highly idiosyncratic. Differences in life history necessarily lead to differences in memorised information. Moreover, it has been repeatedly shown that two people witnessing the same event even a salient and highly memorable event like a car accident may construct dramatically different representations of it, disagreeing not just on their interpretation of it, but in their memory of the basic physical facts [(Loftus, 1979; Neisser, 1982)]. While grammars neutralise the differences between dissimilar experiences, cognition and memory superimpose differences even on common experiences.

---

[20] Relevance Theory gets rid of the concept of "what is said" and introduces the concept of *explicature* which serves to characterize the explicit meaning of an utterance, paralleling the Gricean concept of *implicature* which characterizes implicit meaning (see Sperber & Wilson, 1986/1995, §4.2). The point of the concept of "explicature", then, in contrast with "what is said", is to stress that inferential processes are also needed to derive the explicit meaning of utterances (see also Sperber & Wilson, 2012, §1.2 and §1.3).

> Sperber & Wilson (1986/1995, p.16)

Subjectivity therefore fundamentally challenges the effectiveness of language as a communication system, at least insofar as communication success is characterized as the identicity of messages between interlocutors.

This is also where the encounter between Relevance Theory and literature proves particularly fruitful actually, for as Wilson (2018) tells us:

> According to the code model, what is communicated is a 'message' or 'meaning' which can be rendered as a proposition (or a small set of propositions) and duplicated in the minds of communicator and addressee; however, as literary scholars are well aware, the thoughts communicated by a literary text are often too rich and vague, too complex and subtly interlinked to be treated as a meaning or message of this type. Moreover, according to the code model, communication is a yes–no matter: a thought is either communicated or it is not; however, as literary scholars are again well aware, the interpretation of literary works often calls for some creative input from the reader, so that different readers arrive at different interpretations for which they must share some of the responsibility themselves.
>
> Wilson (2018, pp.186–187)

In literature indeed, we are always ready to acknowledge that *interpretation* plays a central role in that, ultimately:

> [E]ach reader makes use of [a] book in his or her own way
>
> Wilson (2018, p.190)

In fact, as Fish (1980) even tells us that:

> The objectivity of the text is an illusion, and moreover, a dangerous illusion.
>
> Fish (1980, p.43)

Yet, this observation actually applies to *non*-literary work as well, such as this very manuscript, for no matter how much effort I put into carefully selecting each word so as to avoid potential misunderstanding or orient possible interpretations, truth is that I have no definite control over how *you* will make sense of it. No matter how hard I try, I can never constrain the interpretation of my own utterances enough to bypass the subjectivity of interpretation. And the example certainly does not stop at my thesis, for every academic surely shares the experience of having once encountered what they would consider to be an unexpected—maybe even unfair—interpretation of their own work.

Now, it is one thing to say that we have no *control* over how our utterances get interpreted in practice, but the thing is that we also have no comprehensive and explicit *knowledge* of how they get interpreted either. As Wilson (2018) rightly emphasizes indeed, we have no knowledge of all the interpretations that *can* be made of our own utterances and no knowledge of the interpretations that *will* be made, since:

> [T]he addressee [interpreter] must necessarily [. . .] draw some conclusions that the communicator need not necessarily either anticipate or endorse.
>
> Wilson (2018, p.190)

I, for one, certainly have no idea of all the possible interpretations that can be made of my own thesis, and I am actually looking forward to discovering them all and to meeting the various subjectivities that will have produced them. This, after all, is what makes research so exciting, for if interpretation was actually so straightforward and "objective", there would probably be no point in doing research in the first place.

Subjectivity thus fundamentally challenges the perspective of mutual understanding, given how we always make sense of each other's utterances through our individual and potentially idiosyncratic mental spaces. You could think that such considerations are artifactual of literary examples which forget that linguistic communication is more often than not an *interactive* process where interpreters can actually fix and adjust their interpretations through a back-and-forth process according to the feedback they receive—and so ultimately converge to mutual understanding. But remember that subjectivity means both idiosyncrasy *and* privacy of the mental, so that if interpretation is *subjective*, it is also because making sense is a fundamentally *private* process. This is precisely why, when Shannon & Weaver (1949/1964, p.68) imagine some sort of "noise correction" mechanism able to fix discrepancies across encoded and decoded messages so as to guarantee communication success, they appeal to a third-party "observer" distinct from both the transmitter and the receiver (see Figure 1.7). In practice, however, there is no omniscient third-party observer that can guarantee the identicity of messages and interpretation always proceeds *de*

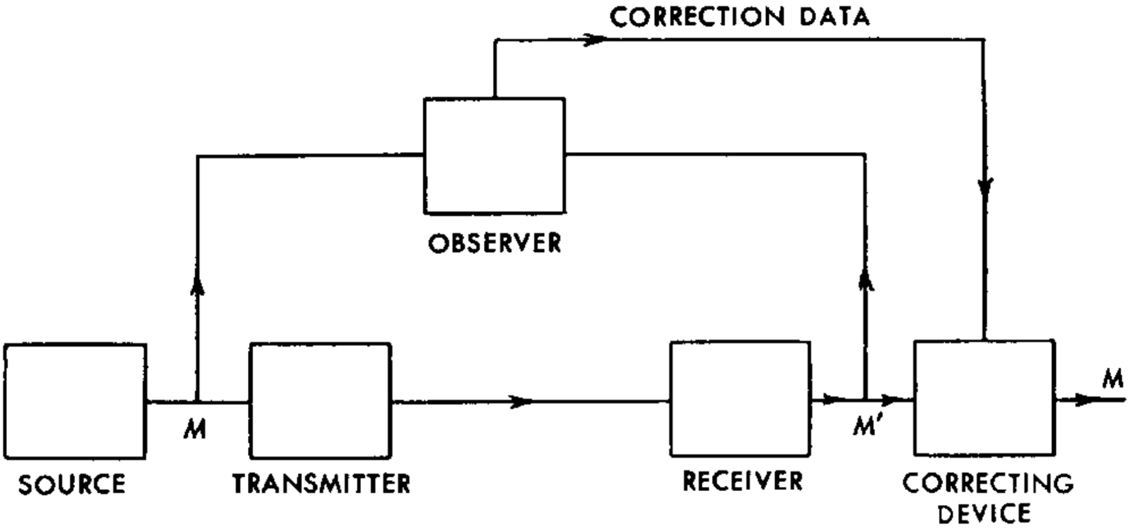


Figure 1.7: Noise correction in the code model of communication. Original diagram from (Shannon & Weaver, 1949/1964, Figure 8, p.68).

facto in full subjectivity. The assumption of mutual understanding thus rests on a critical mismatch between the *public* characterization of communication success and the *private* nature of the interpretation process. In the end, the fact of the matter remains that in conversation—as in any type of human interaction for that matter—we do not have direct access to each other's brain.

1.4.3 Why communication is not signaling

I have acted so far as if *successful* communication could be equated with *effective* communication, and in fact the two notions are very often used interchangeably in practice. Yet, in their seminal work, Shannon & Weaver (1949/1964) actually distinguish the two by separating what they call the "semantic problem" from the "effectiveness problem". In this section, I would like to return to this fundamental distinction to show how crucial it is for our current discussion regarding the communicative function of language.

According to Shannon & Weaver (1949/1964, p.4), the *semantic problem(s)* is "concerned with the identity, or satisfactorily close approximation, in the interpretation of meaning by the receiver, as compared with the intended meaning of the sender" (see Definition 11). The *effectiveness problem(s)*, on the other hand is "concerned with the success with which the meaning conveyed to the receiver leads to the desired conduct on his part" (see Shannon & Weaver, 1949/1964, p.5, and Definition 12). In short: there is effect on the *mind* on one hand and effect on *behavior* (conduct) on the other.

Definition 11 (The semantic problem)
The semantic problem corresponds to the following question: how precisely do the transmitted symbols convey the desired meaning?

Adapted from Shannon & Weaver (1949/1964, p.4)

Definition 12 (The effectiveness problem)
The effectiveness problem corresponds to the following question: how effectively does the received meaning affect conduct in the desired way?

Adapted from Shannon & Weaver (1949/1964, p.4)

The major benefit of this formalization of communicative effectiveness is that it becomes a measurable entity—given how it appeals to observable behavior—yet it is not quite the one we are after. What we are after here indeed is communication success defined as mutual understanding and formally characterized as the identicity of messages—which thus corresponds to the *semantic problem* of Shannon & Weaver (1949/1964) and not to their effectiveness problem. Be that as it may, could we use one (communicative effectiveness) to evaluate the other (mutual understanding)? The answer is not so clear, and Shannon and Weaver themselves tell us that:

> The effectiveness problem is closely interrelated with the semantic problem, and overlaps it in a rather vague way.

Shannon & Weaver (1949/1964, p.5)

First of all, communicative effectiveness does not seem to be a *necessary* condition to attest of mutual understanding. Indeed, ineffectiveness in communication does not necessarily imply the absence of mutual understanding: I could ask you (1.21) and you could very well refuse to pass me the salt for whatever reason (argue that I have had enough already, that you do not want to help me, etc.).

(1.21) Can you pass the salt please?

But refusing to behave the way I wanted you to here would not mean that you necessarily misunderstood what I meant: you could have perfectly derived my communicative intentions and yet refuse to satisfy them nonetheless.

Now, one could argue that this characterization of effectiveness is too restrictive, for there could exist many different types of behaviors that would attest of our mutual understanding independently of your satisfying my intentions—such as your replying (1.22) to my (1.21), for instance.

(1.22) No, you've had enough already.

But the deeper problem with relying on behavior to attest of mutual understanding is that human communication does not always trigger *overt* and *immediate* behavioral feedback from the interpreter. If I talk to you with no other communicative intention than to convey information—such as uttering (1.23)—your processing of that information does not necessarily imply some sort of immediate behavioral response on your part.

(1.23) It is raining outside.

This is precisely the point made by Lewis (1975/1983) when he reflects of his earlier characterization of "signaling games" which originally required interpreters (*hearers*, in his case) to always respond to a received signal by an action (see Lewis, 1969/2002, p.131, for the original specification of that point):

> [S]ignaling was all-too-obviously a special case. There may be no appropriate responsive action for the hearer to perform when the speaker imparts a belief to him. Or the speaker and hearer may disagree about how the hearer ought to act under the supposed circumstances. Or the speaker may not know how the hearer will decide to act; or the hearer may not know that he knows; and so on. The proper hearer's response to consider is *believing*, but that is not ordinarily an action.
>
> Lewis (1975/1983, p.170)

If you think of it, this fundamental consideration is at the core of the Gricean characterization of communication success as intention resolution in Definition 4 for strictly speaking, there is nothing *behavioral* in having someone recognize your intention.

The question, then, is what to do with all those cases that trigger no overt and immediate behavioral feedback in communication: how can we, in such cases, attest of mutual understanding? One possible solution would be to equate success with the absence of explicit communication failure or "conflict" between interlocutors: so long as you do not express or demonstrate that you have not understood me, I will just assume that you have. But that would not be satisfactory either, all the more so as even explicit agreement is no guarantee of mutual understanding—as Shannon & Weaver (1949/1964) themselves actually make quite clear when they tell us that:

> If Mr. Y says "Do you now understand me?" and Mr. X says "Certainly, I do," this is not necessarily a certification that understanding has been achieved.
>
> Shannon & Weaver (1949/1964, p.4)

One could argue that all those cases should prove marginal in human communication overall and that most of the time, communication *does* trigger some sort of overt and immediate behavioral feedback from the interpreter. But precisely, the difference here may not prove so much *quantitatively* significant than *qualitatively* so, as it may be what specifically distinguishes (human) *communication* from (animal) *signaling*. As Reboul (2017) tells us, for instance:

> [W]hile animal signals are paired with responses, linguistic signals, on the whole, are not.
>
> Reboul (2017, p.42)

And as she describes further, there seems to be a fundamental qualitative difference of language use between human and non-human animals:[21]

> [I]t is interesting to note that animals engaged in animal language programmes spontaneously use 'language' only to request things or actions from others (mostly their caretakers) [...] By contrast, humans are able to deploy thoughts that [...] bear on absent targets [and] are not oriented towards action (indeed, the place that fiction—or, on a more anthropological vein, myth—takes in human lives is a good indication of that). Millikan [(2013)] has rightly insisted on the potentiality for strong disinterestedness in human thought: humans, and—as far as we know—only humans, are able to entertain thoughts that do

---

[21] From a different perspective, Tomasello (2019) tells us that:

> Systematic studies have found that over 95 percent of the communicative acts produced by [apes] are some form of imperative [(and the other 5 percent are things like naming games: Greenfield & Savage-Rumbaugh, 1993; Rivas, 2005)]. This is because no matter how they are trained by humans, great apes will not acquire a motive to simply inform others of things or share attitudes with them [(Tomasello, 2008)]. They also do not learn such things as pronouns (for example, he, she, it), which require assessment of common ground. (Tomasello, 2019, p.124)

> not bear on their immediate environment and that, equally, are not oriented towards any action. Hence, the human ability for speculative thought of a non-practical kind.
>
> Reboul (2017, p.101)

In the end, communicative effectiveness framed in behavioral terms appears to be neither *necessary* nor *sufficient* to attest of mutual understanding, and so pretty much useless as it stands (see Table 1.1 from a recap of the arguments).[22] But what if the problem lied not with the notion of communicative effectiveness but with that of mutual understanding? Because saying that communicative effectiveness is not a sufficient condition to attest of mutual understanding also means, by equivalence, that mutual understanding is not a necessary condition to attest of communicative effectiveness (see Equation 1.24).

Question	**Answer**
effectiveness $\overset{?}{\Longleftarrow}$ mutual understanding Is communicative effectiveness a *necessary* condition to attest of mutual understanding?	effectiveness $\not\Longleftarrow$ mutual understanding No: ineffective communication does not necessarily imply misunderstanding. ($\neg$ effectiveness $\not\Longrightarrow$ $\neg$ mutual understanding)
effectiveness $\overset{?}{\Longrightarrow}$ mutual understanding Is communicative effectiveness a *sufficient* condition to attest of mutual understanding?	effectiveness $\not\Longrightarrow$ mutual understanding No: misunderstanding does not necessarily imply ineffective communication. ($\neg$ mutual understanding $\not\Longrightarrow$ $\neg$ effectiveness)

Table 1.1: Why communicative effectiveness is neither *necessary* nor *sufficient* to attest of mutual understanding (effectiveness $\not\Longleftrightarrow$ mutual understanding).

$$\{\text{effectiveness} \not\Longrightarrow \text{mutual understanding}\} \Longleftrightarrow \{\text{mutual understanding} \not\Longleftarrow \text{effectiveness}\} \tag{1.24}$$

So what if our demonstration so far actually captured a crucial intuition about communication, namely, that irrespective of how we chose to characterize communicative effectiveness, the concept of mutual understanding was not necessary in the first place? What if there was no "semantic problem" from the start?

Everyone having experience being immersed in a foreign environment where you do not speak the language can attest that a lack of (mutual) "understanding" does not necessarily prevent effective communication with other speakers. If anything, this example further illustrates that human communication extends much beyond *linguistic* communication proper and that interpretation can often proceed by purely inferential means. Yet,

---

[22] As detailed in Table 1.1, the reasoning used throughout this section relies on *proof by contraposition*, that is, given two propositions P and Q, on the fact that $\{P \Longrightarrow Q\} \Longleftrightarrow \{\neg Q \Longrightarrow \neg P\}$.

there is also an obvious qualitative difference between speaking a foreign language and *not* speaking that language—and being able to effectively communicate with an English speaker by speaking English is clearly not the same thing than being able to effectively communicate with that same speaker while *not* being able to speak English. So if we could be ready to drop the concept of *mutual understanding* given its uselessness in characterizing communicative effectiveness, we might no want to drop the concept of *understanding* altogether, although for that we would need a characterization of communication success that is completely separated from that of mutual understanding. But is that even conceivable? This is the question I would like to turn to now.

1.5 Communication beyond mutual understanding

1.5.1 The presupposition of mutual understanding

The original critique of the code model carried out by Sperber & Wilson (1986/1995) concentrates notably on what they call the "mutual knowledge hypothesis" which is mobilized by proponents of the code model to accommodate the pervasiveness of context in communication (see Sperber & Wilson, 1986/1995, §1.3).

The problem that the mutual knowledge hypothesis is supposed to solve is as follows. Given, first, the *underdetermined* nature of messages which requires every communicated signal to be contextualized in order to be decoded. Given, second, that context is something that interpreters *build* rather than something they "pick up" as if it was made readily available to them. How can one make sure that the context used by the interpreter at decoding time will be the exact same one as the one used by the speaker at encoding time?

Under the framing of the code model, the underdetermination of messages and the subjectivity of the contextualization process pose a definite challenge to the very possibility of communication success, for if communication proceeds by means of a *shared code* between all communicating parties, then there is no alternative but to require context to be strictly identical across interlocutors in order to guarantee mutual understanding: the *transmitted* signal–context pair must necessarily be identical to the *received* signal–context pair indeed if it is to map straightforwardly and identically to the exact same message across interlocutors.

To make sure that they actually rely on the exact same context, interlocutors must then overcome both the idiosyncratic *and* private nature of the contextualization process. To do so, they must not only make sure to rely on contextual knowledge that they *share*, but also on contextual knowledge that they *know* to be shared—that is, on contextual knowledge that is *mutually known*—and it is with this latter point that Sperber & Wilson (1986/1995) take issue.[23] As they argue indeed, keeping track of their mutual knowledge at encoding

---

[23] The terminology used by Sperber & Wilson (1986/1995) is a little confusing because the literature usually distinguishes *mutual* knowledge from *common* knowledge where *mutual knowledge* of p means "everybody knows p" and *common knowledge* of p means "everybody knows that everybody knows p (recursively)" (see Vanderschraaf & Sillari, 2022, for an overview). So what Sperber & Wilson (1986/1995) actually refer to here when they use the term "mutual knowledge" is what is traditionally referred to as

and decoding time requires interlocutors to perform an infinite series of recursive checks across each other's knowledge, which seems psychologically implausible in the standard time frame usually allotted to encoding and decoding in communication (Sperber & Wilson, 1986/1995, p.18).

Now, as Lewis (1969/2002) fundamentally stresses, this "infiniteness" of recursion is a theoretical idealization and is not supposed to have any real psychological bearings:

> Note that this is a chain of implications, not of steps in anyone's actual reasoning. Therefore there is nothing improper about its infinite length.
>
> Lewis (1969/2002, p.53)

Though Schelling (1960/1965, p.54) shows with his "department store example" that this idealization could be a relatively good psychological *approximation* nonetheless: if you and I are looking for each other in a department store, the situation does theoretically require us to perform an infinite number of recursive checks about our respective assumptions— *Where could you possibly think I could think you could think [...] I am?* and *vice versa*—and yet we do seem to be able to manage those situations in our daily lives nonetheless.

Be that as it may, Sperber & Wilson (1986/1995) precisely take issue with the fact that mutual knowledge is an *idealization*—not a reality, in fact, but:

> [...] an ideal people strive for because they [...] want to avoid misunderstanding whenever possible.
>
> Clark & Marshall (1981, pp.18–19)

And even more than with its being an idealization, with its being a *theoretical necessity*:

> Pragmatists have no positive argument that individuals engaging in verbal communication can and do distinguish mutual from non-mutual knowledge. Their only argument is a negative one: if mutual knowledge does not exist in the form required by the code model of verbal communication, then the code model is wrong. Since they see the code model as the only possible explanation of communication, they cling to the mutual-knowledge hypothesis. Instead of adopting the code model, seeing that it commits us to the mutual-knowledge hypothesis, and then having to worry about how this hypothesis can be empirically justified, we want to approach things the other way around. We see the mutual-knowledge hypothesis as untenable. We conclude, therefore, that the code theory must be wrong, and that we had better worry about possible alternatives.
>
> Sperber & Wilson (1986/1995, pp.20–21)

"common knowledge" in the literature. To avoid adding further confusion to the mix, I will stick to their original terminology and use the term "mutual knowledge", but bear in mind that I am using it here to refer to "common knowledge".

I have chosen to begin this section with the critique or the mutual knowledge hypothesis from Sperber & Wilson (1986/1995, §1.3) for the argument as much as for the argumentation itself. Because what the quote from Sperber & Wilson (1986/1995, pp.20–21) above fundamentally stresses is that the mutual knowledge hypothesis is not something that unfolds naturally from careful empirical observation and investigation, but a theoretical necessity and as such an *assumption*—something that actually *needs* to be true in order to satisfy some other theoretical requirement that precedes it in the overall theory; some other *presupposition*, in fact. Here, the presupposition in question is the idea that human communication proceeds by means of a shared code between all communicating parties—a "shared code presupposition" which is the main target of Relevance Theory which argues that human communication is much more inferential than coded and which pushes for abandoning the code model altogether in favor of a more inferential theory of human communication.

Yet, the code model is not reducible to this shared code presupposition alone: as most theories, it rests on a *set* of presuppositions and another one of its foundational assumption is what I call the "presupposition of mutual understanding"—the assumption that communication proceeds by duplicating a message (or "thought") from the mind of a speaker into the mind of an interpreter.[24] It is the very assumption that lies behind the formal characterization of communication success as the strict identicity of encoded and decoded messages in the code model. My purpose here then, and throughout this Section 1.5, is to argue that the exact same type of reasoning can be carried out regarding the presupposition of mutual understanding than with the shared code presupposition. Just like the idea that language proceeds by means of a shared code between communicating parties, the idea that communication consists in mutual understanding is a *presupposition* of the code model—an assumption that is taken for granted but that is actually far from being empirically justified. And just like the shared code presupposition before it, I will argue, it is a presupposition that *can* and *should* be changed.

1.5.2 The requirement of shared meaning

> Do others understand what we say or write? Do we understand them? These are questions not often addressed in language theory. Those professionals who work in language theory—literary theorists, linguists, philosophers of language, communication theorists, semioticians, theorists of rhetoric, discourse analysts, etc.—are more interested in the problem of specifying *what* it is to understand and *how* we understand than in asking *whether* we understand. Apparently, the fact that communicators ordinarily understand each other is a pre-theoretical given, the sine qua non of academic discourse on language, meaning, and interpretation. Consequently, asking whether we understand our fellow communi-

---

[24] In his book *On Writing: A Memoir Of The Craft*, Stephen King answers the question "what is writing?" by saying:

> Telepathy, of course. (King, 2014, p.103)

I am grateful to Jacques Moeschler for pointing me to this reference.

> cators is typically treated as the sort of non-serious question that only a radical sceptic would even consider raising.
>
> Taylor (1992, p.3)

Any presupposition of a scientific theory can generate in turn a set of assumptions that crucially depend on it and are mobilized to make the axiomatic system hold together in a consistent fashion. The presupposition of mutual understanding is no exception, as it generates what I call "the assumption of shared meaning"—which is part of what Taylor (1992) calls the "dogma of the double conformity" (Definition 13) and which Locke (1690/2004) originally considered to be one of our fundamentally mistaken beliefs about language (see Taylor, 1992, pp.27–46, for details).[25]

> **Definition 13 (Dogma of the double conformity)**
> Our actions are formed on the presupposition that the ideas we signify by our words are the same ideas as others signify by those words and that our ideas accurately represent things.
>
> Adapted from Taylor (1992, pp.28–29)

The articulation between the presupposition of mutual understanding and the assumption of shared meaning is actually made explicit by Locke (1690/2004) when he tells us that:

> And hence it is, that men are so forward to suppose, that the abstract ideas they have in their minds, are such as agree to the things existing without them, to which they are referred; and are the same also, to which the names they give them, do, by the use and propriety of that language belong. For without this *double conformity* of their *ideas*, they find they should both think amiss of things in themselves, and talk of them unintelligibly to others.
>
> Locke (1690/2004, II.xxxii.8)

And indeed, as Taylor (1992) explains:

> Locke takes our mistaken belief in intersubjective conformity [shared meaning] to be derived from our taken-for-granted belief that communicators ordinarily understand each other [mutual understanding]. In other words, the reason people believe that the idea they attach to a word W must be the same as the idea other people attach to W is that, if this were not the case, they would fail to communicate with others whenever they used W. But this conclusion seems to clash with commonsense: that is, people ordinarily take themselves to communicate successfully with their interlocutors. Thus, according to Locke, because they believe themselves to be communicating successfully whenever they use words such as W, they therefore believe that other people attach the same ideas to words that they do.

---

[25] For an overview of Locke's philosophy of language, see also (Powell, 2021; Hill, 2021).

> Taylor (1992, p.29)

In other words, the assumption of shared meaning is a theoretical requirement that derives from the presupposition of mutual understanding itself: it *needs* to be true in order to satisfy that presupposition. In fact, the argument seems rather intuitive, for how could we understand one another indeed if we did not share meaning in the first place?

> [P]eople do not associate any old concept to a word. Instead, they learn through socialization which concepts go with which words. So, as a child, you learned that dog refers to a certain kind of animal. If you first developed the hypothesis that dog refers to any four-legged mammal, you would soon find yourself miscommunicating with people. They would not understand you when you referred to a sheep as dog, and you would not understand them when they said that all dogs bark, and so on. Thus, there is a social process of converging on meaning that is an important (and neglected) aspect of language [. . .]
>
> Murphy (2002, p.391)

Yet, for Locke (1690/2004), this assumption of shared meaning is actually misguided:

> [M]en talk to one another, and dispute in words, whose meaning is not agreed between them, out of a mistake, that the signification of common words, are certainly established, and the precise ideas, they stand for, perfectly known; and that it is a shame to be ignorant of them. Both which suppositions are false: no names of complex ideas having so settled determined significations, that they are constantly used for the same precise ideas.
>
> Locke (1690/2004, III.xi.25)

However, his position seems to contradict one of our most basic intuition about language and meaning, namely that “the meanings of individual words are by and large shared and stable” (Jackman, 2020, p.1). So do we, in fact, share meaning with one another?

Answering that question is actually far from trivial, but what is interesting about it is that it has very practical consequences. In the field of computer science, for instance, the assumption of shared meaning originally lead computer scientists in the 1980s to try and design text-based (verb-mediated) user interfaces to personal computers before settling for the graphical interfaces we are now most widely accustomed to. What happened which made them give up on those text-based interfaces is that they faced the extent of the subjectivity of meaning: users turned out never to use quite the same words to refer to the same functions/concepts, or to think of the same functions/concepts when using the same words, rendering those text-based interfaces unusable in practice (see Ludlow, 2014, pp.73–75 for the original argument). To be more specific, Furnas et al. (1987) show that the probability that any two person would use the same word to refer to the same function ranges between 7% and 18% only, and that when they *do* use the same word to refer to some specific function, the probability that they have the exact same function in mind is actually only about 15% (see also Brennan, 1998, on that point).

Marconi (1997, ch.3) tells us that we should expect word meaning to vary widely across speakers given how language should reflect what Putnam (1975) calls the "division of linguistic labor", itself deriving from the division of *non*-linguistic labor, so that difference in expertise should translate as difference in *conceptual* knowledge and ultimately as difference in *lexical* knowledge. But he also tells us that a non-negligible part of the lexicon should remain immune to such knowledge effects, especially for "concrete" and common nouns. After all, everybody knows what a *spoon* is, for instance, so it does not seem unreasonable to assume that we should all share the same meaning for such words. Yet Labov (1973) actually shows in his seminal work on the semantics of tableware items that even for such concrete words as *mug*, *cup*, *bowl* and *vase* denotation could vary widely from one person to the next, especially if you started modifying specific properties of the object in question such as *width*, *depth*, *content* or even *presence or absence of a handle*.[26]

Truth is that most of the time we take shared meaning for granted and do not really question whether or not it exists. Just like we do not really question *whether* people do understand one another in communication, we do not really question *whether* shared meaning exists, but focus rather on *what* it is made of and *how* it should be characterized in practice. Confronted with the subjectivity of meaning, some propose to distinguish public "concepts" from private "conceptions" (Burge, 1989/2007; Millikan, 2000)—or objective "thoughts" from subjective "ideas" (Frege, 1897/1979)—so as to accommodate subjectivity all the while not renouncing the assumption of shared meaning by separating the purportedly shared, public and objective aspects of meaning from the idiosyncratic, private and subjective aspects of our beliefs and knowledge about the world. In some sense, the suggestion does not seem quite unreasonable, for isn't there something "objective" about the world that persists irrespective of what we believe or know it to be?

> When you and I look at the sun, my act of perception is distinct from your act of perception, but this should not lead us to conclude that the object perceived is also different for us. Indeed not, we both perceive numerically the *same* object: the sun.
>
> Taylor (1992, p.102)

And whether I believe a cat to be typically grey or striped, big or small, cute or scary, and whether I disagree with you on the matter or not, isn't it possible for me to understand you regardless when you tell me that "there is a cat on the mat"? To even say that we "disagree"

---

[26] Malt (2020) tells us more or less the same thing today, when she says that:

> For large samples of pictures of dishwares, household containers, and instances of locomotion, the mean agreement on the name produced to a stimulus by healthy native speakers has been found to range from 52% to 83% depending on language and domain [(see Footnote 5 in Malt et al., 2015, for details)]. (Malt, 2020, p.244)

about what a cat is, don't we need to agree on what we disagree in the first place? (see Jackman, 2020, especially §3.2). As Chomsky (2000) rightly emphasizes, however:

> It seems that other cognitive systems—in particular our system of beliefs concerning things in the world and their behaviour—play an essential part in our judgements of meaning and reference, in an extremely intricate manner, and it is not at all clear that much will remain if we try to separate the purely linguistic components of what in informal usage or even in technical discussions we call the "meaning of a linguistic expression". I doubt that one can separate semantic representation from beliefs and knowledge about the world.
>
> Chomsky (1979/2007, p.142)

In fact, most of those techniques attempting to objectivize shared meaning boil down to marginalizing subjectivity one way or another—either by saying that it does not really exist, or that it does not actually matter. But the practice is untenable, for as Pelletier (2017) fundamentally summarizes:[27]

> [D]ifferent subjects give individually different results on the many tasks about meaning that have been administered over the decades in cognitive psychology.
>
> Pelletier (2017, p.74)

In the end, however, shared meaning may not matter much for mutual understanding actually, for it does not seem to constitute a necessary (let alone sufficient) condition to guarantee communication success:

> The fact that a public word exists, and is successfully used in communication, does not make it safe to assume that it encodes the same concept for all successful users; and in any case, the concept communicated will only occasionally be the same as the one encoded. Communication can succeed, despite possible

---

[27] For a recent account, see (Lupyan et al., 2022; Duan & Lupyan, 2023). Else, to illustrate Pelletier's point concretely: psychological experiments on lexical meaning—which typically ask subjects to grade lists of word pairs on a ten-point scale, or triangular arrays of words by choosing among a pair of word the most similar to a referent word—exhibit mixed levels of agreement across subjects: from 44% to 63% on word pairs and from 45% to 66% on triangular arrays depending on the categories being tested for (Hutchinson & Lockhead, 1977). You could think that those results are artifactual of experimental setups that study word meaning in isolation and separated from (sentential) context, but truth is that lexical substitution tasks do not fare much better: agreement ratings across subjects settle for a low 28% in (McCarthy & Navigli, 2009) and as low as 19% in (Kremer et al., 2014) and 16% in (Sinha & Mihalcea, 2014). In (Federmeier & Kutas, 1999) expectations regarding missing words in a clozed test average at 74%, while ranging from 17% to 100% depending on tested items. Even if we took "shared meaning" to only apply to some specific aspects of lexical or conceptual structure, such as "cores" (e.g. Barsalou, 1982, 1989) or "prototypes" (e.g. Rosch, 1973, 1975, 1978) the question would be far from settled. Indeed Barsalou (1987) shows in a large-scale replication study that agreement on prototypes actually ranges between 45% and 50% across subjects, significantly below the original 90% reported by (Rosch, 1975). And in the end, both cognitive science and linguistics alike tell us that concepts may not have cores or stable and/or shared structure, but may rather be "ad hoc" and constructed on-the-fly as need be (e.g. Casasanto & Lupyan, 2015; Carston, 2019).

> semantic discrepancies, as long as the word used in a given situation points the hearer in the direction intended by the speaker. [...] More generally, it does not much matter whether or not a word linguistically encodes a full-fledged concept, and, if so, whether it encodes the same concept for both speaker and hearer. Even if it does, comprehension is not guaranteed. Even if it does not, comprehension need not be impaired.
>
> Sperber & Wilson (1998, p.200)

An argument that Chomsky (2000) extends to the notion of "shared language" as well:

> It is often argued that such notions as common "public language" or "public meanings" are required to explain the possibility of communication [...] Thus, if Peter and Mary do not have a "shared language," with "shared meanings" and "shared reference," then how can Peter understand what Mary says? (Interestingly, no one draws the analogous conclusion about "public pronunciation.") [...] But these views are not well founded. Successful communication between Peter and Mary does not entail the existence of shared meanings or shared pronunciations in a public language [...] any more than physical resemblance between Peter and Mary entails the existence of a public form that they share.
>
> Chomsky (2000, p.30)

So maybe we do not need the assumption of shared meaning in our theory of language and communication, as we do seem to be able to account for our intuitions of "communication success" without it. But what if we did not need the presupposition of mutual understanding either, actually? What if we could separate the question of communication success from that of mutual understanding, and retain the former while getting rid of the latter? What if we did not understand each other after all?

1.5.3 The subjectivist approach to communication

When I look at my feet, they tell me that the Earth is flat. This seems intuitive, common sense. And yet, it is not (Asimov, 1972). But there is something about the Earth, something about its *size* especially, that makes it possible for me to foster that illusion. What if the same was true of language and mutual understanding? What if there was something about language that made it possible to foster what would prove to be nothing more than an *illusion* of mutual understanding?

We have the impression that we largely share meaning with others, and yet when we look at it closer we see that we overestimate how much we actually do (Martí et al., 2019, 2023). We have the impression that our interlocutors understand us in general, and yet when we look at it closer we see that we overestimate how much they actually do (Keysar & Henly, 2002; Lau et al., 2022). Truth is that we are probably much more egocentric in communication than we think: we always presume that our interlocutors are "like us" by default, and only take their subjectivity into account when we really have to—that is, only when communication actually breaks down (Keysar et al., 1998; Nickerson, 1999). The problem with communication, however, is that it does not necessarily break down because of a misunderstanding, and that effective communication is not a guarantee of perfect mutual understanding either (see Section 1.4.3). In fact, Gärdenfors (2014) makes clear that what we need to make communication work in general is not so much the *confirmation* that we understand one another than the mere *belief* that we do:

> Just as bargainers shake hands after reaching agreement on the terms of a contract, so speakers reach a point at which both believe they have understood what they are talking about. Of course, they may actually mean different things, just as the bargainers might interpret the terms of the contract differently. It is enough that, in a given moment and context, speakers reach a point at which they *believe* there is mutual understanding.
>
> Gärdenfors (2014, p.130)

The more fundamental problem with communication seems to be that we often confuse our intuitions of communicative effectiveness with our intuitions of mutual understanding.[28]

There is also something about language, meaning and communication in general, that makes it possible for us to maintain our illusions about shared meaning and mutual understanding. It is precisely because of *underdetermination*, for instance, and because the words that we utter do not even being to convey the richness of our underlying conceptual representations, that we can maintain the illusion that meaning is shared. And it is precisely because *communication is not signaling* and because human communication does not systematically trigger overt and immediate behavior feedback from the interpreters that we can maintain the illusion that we understand one another.

In any case, abandoning the presupposition of mutual understanding is not that easy, as it appears to pertain to the very *nature* of communication itself. What would be left of communication indeed without mutual understanding? Whichever way you decide to flip the problem, the two notions appears to be fundamentally intertwined, to the extent that it seems impossible to conceive one without the other.

So what I propose to do here is to proceed step-by-step, and to clarify a few points. First of all, abandoning the presupposition of mutual understanding does not mean that we can no longer talk about communication success. What it means, however, is that we no longer need to characterize communication success in objective terms—as we usually do with the notions of identicity or sufficient similarity between messages, for instance—as we no longer need to even compare produced and interpreted messages with one another and no longer need to talk about "correct" or "incorrect" interpretation either.[29] As such, communication without mutual understanding entails a fundamental change of paradigm

---

[28] I will return to that point more specifically in Section 2.3.1.

[29] Which does not mean that we no longer need to talk about communicative effectiveness in objective terms, however, but precisely that is a separate matter.

in how we approach communication itself, since it is now best characterized as the *subjective*—rather than *inter*subjective (and hence *objective*)—coordination of minds.[30]

Second, abandoning the presupposition of mutual understanding and the assumption of shared meaning along with it does not mean that *anything goes* as far as meaning and interpretation are concerned. For indeed, subjectivism is too often cast away as a form of radical meaning relativism leading to potentially absurd situations as the one popularized by Lewis Carroll (1871/1896) in the following passage of *Through the Looking-Glass*:

> [Humpty Dumpty:] "And only *one* for birthday presents, you know. There's glory for you!"
>
> "I don't know what you mean by 'glory,'" Alice said.
>
> Humpty Dumpty smiled contemptuously. "Of course you don't—till I tell you. I meant 'there's a nice knock-down argument for you!'"
>
> "But 'glory' doesn't mean 'a nice knock-down argument,'" Alice objected.
>
> "When *I* use a word," Humpty Dumpty said in rather a scornful tone, "it means just what I choose it to mean—neither more nor less."
>
> Carroll (1871/1896, pp.106–107)

An argument which Winograd (1972) otherwise frames quite clearly as follows:

> If the "meaning" of an utterance can only be described in terms of its effects on a particular understander with a particular history, how do we talk about inter-subjective meaning at all? Since no two people have identical histories, and since any aspect of cognitive structure can potentially have an effect on the processing triggered by a particular utterance, there is a different meaning for every hearer. There is no objective "right meaning"—only a meaning for a particular person at a particular moment in a particular situation. Carried to an extreme, if you interpret my use of the word "dog" as referring to eggplant Parmesan, what allows me to argue that you are wrong? We want to

---

[30] For an overview of the concept of "intersubjectivity" in language and discourse, see (Linell, 2017). For an overview of the concept in the context of linguistic anthropology, see (Enfield et al., 2014, Part III). In (Enfield et al., 2014), Sidnell (2014) explicitly connects the concept of intersubjectivity with that of mutual understanding in telling us that:

> The structures of talk-in-interaction provide for a form of intersubjectivity or mutual understanding that is distinctive in the animal kingdom. (Sidnell, 2014, p.364)

Linell (2017) also tells us that:

> [P]reliminary accounts of intersubjectivity often home in on shared understandings among participants in the situation or community involved. Many scholars have taken intersubjectivity as a necessary assumption or a definitional point-of-departure for successful communication. (Linell, 2017, p.109)

Else see (Taylor & Cameron, 1987; Zlatev et al., 2008; Froese, 2018; De Jaegher, 2018; Conrad, 2022) and references therein for relevant references on the concept of intersubjectivity. I will return to the distinction between *intersubjective* and *subjective* coordination in more details in Chapter 2.

> understand meaning in a way that makes sense of the fact that you and I may not have identical (or even mutually consistent) understandings of "democracy" or "system," but we cannot ignore the common sense intuition that there are broad areas of obvious agreement.
>
> Winograd (1980, p.225)

But precisely, as Winograd suggests, the example here is "carried to an extreme". The fact of the matter is that there is ample leeway between my using *dog* to refer to your *eggplant parmesan* and you and I having perfectly identical (or even "similar", for that matter) concepts of *dog*.[31] Language leaves out infinite nuance in between shared meaning and overt conflict, and it is within this infinite nuance that subjectivism steps in.

To give you a sense of my subjectivist intuitions, let me use a mathematical example. Consider the problem in (1.25), formalized here using some arbitrary mathematical notation but which can otherwise be translated into plain English as "find all natural numbers that are strictly greater than three".

$$X = \{x : x \in \mathbb{N},\ x > 3\} \tag{1.25}$$

This problem could be conceived as some sort of *constraint satisfaction problem*: in a sense, the solutions we are looking for must satisfy multiple "constraints" (e.g. be a natural number; be strictly greater than three) and this problem, as you will have noticed, admits more than one solution. To be more specific, it actually admits an *infinite* number of solutions. Yet, does it mean that "anything goes"? Obviously not: every mathematical object that does *not* satisfy all the specified constraints is not a solution in X. The natural number 2, for instance, is not a solution: $x = 2 \notin X$. Now, what does the argument from Winograd (1980) actually translates to in this case? He tells us that if we do not agree on the "constraints" (on meaning) there will be conflict: we will not be solving the same problem. To that, I very much agree. But he also tells us something else: that if we agree on the constraints on meaning, then we necessarily need to agree on meaning *itself*. In other words: meaning (and concept-to-word mapping in this case) is necessarily a closed-problem, if not actually a *puzzle*, that is, a problem that admits a *single* solution. And yet there is no reason *a priori* to consider that meaning should be a closed problem or a puzzle. Why should it be? What prevents us from approaching meaning, language and communication in general, as a potentially unbounded constraint satisfaction problem like (1.25)? More interesting now about the constraint satisfaction problem defined above is that it does not entail any form of intrinsic comparison across solutions—nothing makes 4 a "better" solution to (1.25) than 42, for instance. This is very much the crucial point I wish to make here: by framing the problem of subjective coordination as a *constraint satisfaction problem*, we can readily account for the fact communication may very well

---

[31] With Winograd as with Humpty Dumpty, the problem does not arise because interlocutors assign *different* concepts to the same label, but because they assign *conflicting* concepts to the same label—which makes all the differences.

tolerate a multiplicity of possible interpretations, all the while getting rid of the objectivization of communication success in a way that requires us to talk about *objective* (and hence *normative*) success or failure of interpretation, and so to intrinsically *compare* those interpretations with one another. In subjective coordination, there is no such thing as a “better” or “worse” interpretation of a communicated signal. This way, we no longer need to treat subjectivity as a *problem*. Subjective coordination, as we will see in the next Chapter 2, will not function *despite* or *against* subjectivity but pretty much *around* it.

When I look at my feet, they tell me the Earth is flat. This seems intuitive, common sense. Then I go visit my friends, those who leave by the sea. They tell me the Earth *cannot* be flat, and that it is most likely *round*. For why would boats always seem to sink deeper and deeper into the sea as they sail further and further away from the shore otherwise? “But how can it look so flat to me, then?”, I ask. So together, we refine our theory of the round Earth and conclude that the Earth is not only round but *big*, very very big—much bigger than any of us, in fact—which is precisely why it can look so flat to me when I look at my feet. In doing so, we need not contradict anyone’s intuitions, we need not negate anyone’s subjectivity, or argue that some are necessarily more “correct” than others. We can actually accommodate them all within a single theory of the big round Earth by asking a simple question: what is it exactly that we all have the intuition *of*? and understand that, in my case, what I have the intuition of when I look at my feet is not so much the *shape* of the Earth than its *size*. So will I, in the following Chapter 2, try not to negate or discard our common intuitions about communication success but rather try to accommodate them all within Subjective Coordination Theory by asking what it is exactly that they are intuitions of. All this to say that in language and communication, as in science in general, subjectivity need not be seen as a *problem* in itself. Problems only start when we *stop* at our own subjectivity.

Chapter

2

Communication as subjective coordination

2.1 Summary

In this chapter, I introduce Subjective Coordination Theory—my proposal for a psychological theory of communication that does not rest on the presupposition of mutual understanding introduced in Chapter 1 and, as such, dispenses itself from the necessity of shared meaning, shared language and objective communication success.

I begin by detailing how it distinguishes itself from other psychological approaches to communication by being *non-normative*, that is, by not framing the problem of interpretation as having "correct" or "incorrect" solutions (§2.2.1). I then proceed to specify its core axioms (§2.2.2) and how I propose to operationalize them in practice by relying on the mathematical concepts of *vectors* and *projections* (§2.2.3). I detail how Subjective Coordination Theory constitutes a theory of *acommensurability* in communication which posits that meaning, language and interpretation *need not* be compared across interlocutors in order to characterize communication success, and treats them all as *dynamic* entities, constantly changing as functions of occasional context and idiosyncratic experience (§2.2.4).

I argue that our intuitions about communication success are not intuitions of mutual understanding but intuitions of the *smoothness* of our communicative interactions, that is, of how "easy" or "difficult" it is for each of us to make sense of respective communicated signals (§2.3.1). Using a computational simulation, I demonstrate how Subjective Coordination Theory can account for our intuitions about communication success, and notably for the *virtuous circle of communication*—the intuition that the more we communicate, the easier it becomes for us to communicate successfully (§2.3.2). I conclude by pondering on the evolutionary benefits of subjectivity for collective intelligence and introduce the hypothesis that language could have evolved for the purpose of subjective coordination (§2.3.3).

I then return to the original question of the communicative function of language introduced in Chapter 1 so as to discuss, this time, the proper characterization of the notion of *function* (§2.4.1). I conclude this chapter on some ethical considerations regarding the field of evolutionary linguistics at large, and on what I take to be the necessity to leave out moral considerations in our scientific characterization of the notion of "human nature" (§2.4.2).

2.2 Subjective Coordination Theory

The previous Chapter 1 was a deconstruction chapter. I began by questioning the *efficiency* of language as a communication system—all the way down to its *effectiveness*—and ended up clarifying the theoretical status of mutual understanding. I argued that it constitutes a *presupposition* or an *axiom* in standard theories of communication, and that it could and should be changed. This Chapter 2 intends to be a reconstruction chapter. I will introduce *Subjective Coordination Theory*—my proposal for a psychological theory of communication that does not rest on the presupposition of mutual understanding and, as such, dispenses itself from the necessity of shared meaning, shared language and objective communication success. Before introducing its axioms more formally in Section 2.2.2, let me begin here by discussing first how Subjective Coordination Theory fundamentally departs from other psychological theories of communication, such as Relevance Theory (Sperber & Wilson, 1986/1995) discussed throughout Chapter 1.

2.2.1 A non-normative subjectivism

Subjective Coordination Theory follows what I called in Section 1.5.3 a "subjectivist" approach to communication. I borrow the terminology from Pelletier (2017) who uses it to distinguish, on one side, the approach to meaning followed by formal semanticists and philosophers of language and on the other side, the approach to meaning followed by cognitive scientists and linguists:

> There is a similarity [. . .] between the philosophers of language and the formal semanticists. They both think of "meaning" as being *outside* of speakers, as *public*, as being *in common*. Despite whatever other differences there may be between and amongst them, I therefore will call this merged group *The Objectivists*. [footnote omitted]
>
> There is a different similarity [. . .] between the cognitive psychologists and the cognitive linguists. They both think of "meaning" as something that is *inside* speakers, *private*, as being *personal*. Despite whatever other differences there may be between and amongst them, I therefore will call this merged group *The Subjectivists*.
>
> Pelletier (2017, p.51)

Being a "subjectivist" in my terminology—just like in Pelletier's—means first and foremost approaching communication as a *psychological* phenomenon, that is, considering that "meaning", "language" or "interpretation" are mental processes or entities that take place insider interlocutors' heads, and as such are essentially *private* and potentially *idiosyncratic*, hence *subjective*. Where I depart fundamentally from standard psychological approaches to communication, however, is that my subjectivism does not mobilize any concept of “similarity” whatsoever, and need not do so precisely because it does not rest on the presupposition of mutual understanding. Indeed, what is important to understand about the various approaches questioning the requirement of shared meaning mentioned throughout Section 1.5.2—as illustrated, for instance, by the quote from (Chomsky, 2000, p.30) or the quote from (Sperber & Wilson, 1998, p.200)—is that they mainly target the requirement of *public* meaning or *public* language but do not necessarily advocate for abandoning any notion of “sharedness” in meaning or language whatsoever.[32] Their idea is mainly to replace this requirement of (publicly) shared meaning, shared language or shared interpretation with a more flexible concept of *similarity*.[33] Relevance Theory, for instance, starts precisely on those terms:

> Communication is a process involving two information-processing devices. One device modifies the physical environment of the other. As a result, the second device constructs representations **similar** to representations already stored in the first device. Oral communication, for instance, is a modification by the speaker of the hearer’s acoustic environment, as a result of which the hearer entertains thoughts **similar** to the speaker’s own. The study of communication raises two major questions: first, what is communicated, and second, how is communication achieved?
>
> Sperber & Wilson (1986/1995, p.1, my emphasis in bold)

As a result, one of its main focus has been to argue against the *strict* identicity of (precise) messages of the code model, and to advocate for a more flexible approach to communication success capable of talking about *degrees* of mutual understanding. Reflecting on what they take to be the failure of the code model and the semiotics enterprise at large, Sperber & Wilson (1986/1995) tell us indeed that:

---

[32] Remember how, for instance, Chomsky (2000) stresses that the “physical resemblance between Peter and Mary” did not entail “the existence of a public form that they share” (Chomsky, 2000, p.30) and so how he still appeals to a notion of “resemblance”.

[33] Some could even say that all they do is merely displace the problem of shared meaning, shared language or shared interpretation to that of similarity. This is precisely what Pelletier (2017) takes issue with, for instance:

> It is tempting to say, and many Subjectivists have, [footnote omitted] that communication does not require “absolute identity” of the corresponding mental items but only that they be “similar enough”. But this can’t really be made out in a non-circular way, it seems to me. The hypothesis that two different minds are “similar enough” with respect to their understanding of some term—say, ‘democracy’—has no empirical content other than simply some antecedent belief that the two people *are* understanding what each other says when they talk, despite their differences in acquisition of the relevant concepts. But that was precisely the (alleged) fact that the “similarity” was hypothesized to explain. The very nature of the Subjectivist notion of concept makes it impossible to give an *independent* criterion of “similar enough”. Contrast this with the Objectivist claim that there literally is something—something *in reality*—that the participants both grasp. (Pelletier, 2017, p.53)

> This failure is instructive. What a better understanding of myth, literature, ritual, etc., has shown is that these cultural phenomena do not, in general, serve to convey precise and predictable messages. They focus the attention of the audience in certain directions; they help to impose some structure on experience. To that extent, some similarity of representations between the artists or performers and the audience, and hence some degree of communication, is achieved. However, this is a long way from the identity of representations which coded communication is designed to guarantee. It is not clear how the type of communication involved in these cases could be explained in terms of the code model at all.
>
> Sperber & Wilson (1986/1995, p.8)

Now, Relevance Theory certainly does share a certain "Lockean skepticism" about mutual understanding, at least to the extent that it does not take it for granted:

> [F]ailures in communication are to be expected: what is mysterious and requires explanation is not failure but success.
>
> Sperber & Wilson (1986/1995, p.45)

But this skepticism does not go as far as to call into question the presupposition of mutual understanding altogether. As a matter of fact, its ultimate goal could even be said to be that of *rescuing* mutual understanding, and in particular, rescuing it from the critical challenge posed by *underdetermination*, which it proposes to achieve by supplementing what it takes to be an incomplete code model with an inferential device. For Relevance Theory indeed, if mutual understanding is, in fact, genuinely possible between interlocutors, it is because interpreters rely on their *reasoning* abilities to overcome the underdetermination of communicated messages, which enables them to successfully recover the intended interpretation—an intuition that can actually be traced back to Locke (1690/2004) himself (see Taylor, 1992, pp.119–122, for details).[34]

All in all, Relevance Theory does remain well within the paradigm of mutual understanding, and in doing so it ascribes to a particular form of subjectivism which I call "normative subjectivism". Normative subjectivism is characterized by the fact that it *must* objectivize communication success so as to satisfy the presupposition of mutual understanding, despite its considering "interpretation", "meaning" and potentially "language" itself as subjective entities. Its essential feature then is that it frames communication success in *normative* terms, by considering that there are "correct" and "incorrect" interpretations. Indeed, in Relevance Theory, we do find explicitly mentioned that:

> According to relevance theory, the **correct interpretation** of an ostensive stimulus is the first accessible interpretation consistent with the principle of relevance.

---

[34] Though note that, given what I have just said, "mutual understanding" for Relevance Theory would precisely not be framed *only* as the strict identicity of interpretations across interlocutors but would also encompass the relative similarity or partial overlap between those interpretations.

Sperber & Wilson (1986/1995, p.178, my emphasis in bold)

In contrast, Subjective Coordination Theory adopts what I call a "*non*-normative subjectivism", in that it does not need to objectivize communication success since it does not ascribe to the presupposition of mutual understanding in the first place: in Subjective Coordination Theory, there is not "correct" or "incorrect" interpretation, and no "better" or "worse" interpretation either.

However, saying that we need not talk about "correct" or "incorrect" interpretation does not mean that we can no longer talk about "success" or "failure" in communication. It only means that success or failure is framed solely in *subjective* (and biological) terms—in practice, as cognitive limitations of the interpreter, as we will see in Section 2.3.2.[35] Subjective Coordination Theory does not entail that we can no longer talk about "similarity" either: it merely changes its theoretical status, by making it *optional* rather than *necessary*. As I will detail later on in Section 2.2.4, Subjective Coordination Theory is what I call a "theory of acommensurability" in communication: a theory that tells us that language, meaning and interpretation in general *need not* be compared, not that they *cannot* be compared. This nuance has its importance, for it does not tell us that shared meaning, shared language or shared interpretation do not exist, or that they cannot be characterized one way or another—be it with similarity or otherwise. It merely tells us that those theoretical devices no longer constitute formal requirements in our theory of communication, all of which unfolds directly from abandoning the presupposition of mutual understanding.

In Subjective Coordination Theory, there is no correctness of interpretation—let alone *degrees* of correctness—and no requirement of "sameness" across interlocutors either, be it of meaning, or language. This is very much of an epistemological commitment. For whether or not you did manage to "correctly" interpret my utterance, my point is: *something* happened. What exactly is that "something"? This is precisely what Subjective Coordination Theory intends to find out.

2.2.2 The five axioms of Subjective Coordination Theory

In Section 1.5.1, I insisted on the fact that every scientific theory rests upon a set of *presuppositions* (or *axioms*) that are taken for granted and considered true by definition. By saying that they are true "by definition", I do not mean to imply that they cannot be grounded in intuition or experience—or that they cannot be justified one way or another—but merely that they have a distinct theoretical status than, say, *predictions* (or *propositions*, in the mathematical sense of the term) which are themselves true or false "by demonstration", that is, by observation or derivation of (empirical) evidence in accordance with the original set of presuppositions or axioms. In this section, I would like to introduce the set of axioms that ground Subjective Coordination Theory, and detail whenever possible the intuitions or justifications that motivated them in the first place.

---

[35] Communication will fail if and only if an interpreter is unable to make sense of a communicated signal, that is, if they are unable to find or construct an interpretation of the communicated signal within their conceptual space.

Subjective Coordination Theory rests on five axioms: the Axiom of Productivity (Axiom 1), the Axiom of Atomic Compositionality (Axiom 2), the Axiom of Underdetermination (Axiom 3), the Axiom of Making Sense (Axiom 4) and the Axiom of Least Effort (Axiom 5).

Axiom 1 (Productivity)
Human cognition is unbounded: it can theoretically entertain an infinite number of thoughts.

Axiom 2 (Atomic Compositionality)
Every thought can be decomposed as a finite combination of conceptual atoms called the primitives.

Axiom 3 (Underdetermination)
Any overt behavior or external stimulus underdetermines its corresponding mental representation.

Axiom 4 (Making Sense)
Human cognition strives to process external stimuli.

Axiom 5 (Least Effort)
Human cognition strives to minimize processing cost.

The Axiom of Productivity (Axiom 1) and the Axiom of Atomic Compositionality (Axiom 2) characterize what is traditionally referred to as the property of *discrete infinity*—the foundational assumption that human cognition makes “infinite [use] of finite means”, as per the famous quote of Humboldt (1836/1988):

> [L]anguage is quite peculiarly confronted by an unending and truly boundless domain, the essence of all that can be thought. It must therefore make infinite employment of finite means, and is able to do so through the power which produces identity of language and thought.
>
> Humboldt (1836/1988, p.91)

Note that Subjective Coordination Theory makes no particular commitment as to what constitutes the nature of “thoughts”, other than the fact that they are mental phenomena taking place in the mind of an individual.[36] Notably, it does *not* assume that thoughts are necessarily distinct from “emotions” or “sensations”, or that they require “conscious awareness” or are the sole product of conscious psychological experience. In what follows, I will refer to “thoughts” and “concepts” interchangeably, and make no particular commitment as to the nature of concepts either, other than the fact that they are mental representations.[37]

---

[36] Which certainly does presuppose the existence of something called “the mind”, but this is something I will take for granted.

[37] Not differentiating thoughts and concepts will make it easier to formalize concept composition as a *closed* mathematical operation over the set of concepts C later on in Section 2.2.3. Framed this way, the

If the Axiom of Productivity (Axiom 1) focuses on the unboundedness of human cognition, the Axiom of Atomic Compositionality (Axiom 2) focuses on the "finite means" that make this unboundedness possible.[38] In Subjective Coordination Theory, every though or concept can be *decomposed* as a finite combination of primitives, and can be so because it is always *produced* compositionally.[39] Focusing on the decomposition rather than production aspect of compositionality in Axiom 2 merely serves to emphasize that thoughts and concepts need not systematically be composed "from the ground up" (that is, all the way from primitives) but can be composed from a set of *pre-existing* concepts, themselves already composed from those primitives or from other pre-existing concepts. What remains invariant, however, is that all thoughts and concepts are always decomposable as a finite combination of primitives, even when the corresponding combination process remains intractable (i.e. when we do not know which pre-existing concepts were combined to form the conceptual representation at hand). The set of "primitives" or "conceptual atoms", noted as P, refers to the "building blocks" of human cognition that are assumed to be biologically shared by all human beings by virtue of them being humans. Combined with the compositional operator, noted as $+$, which enables the combination of those building blocks together, they constitute the formal characterization of human cognition. In Subjective Coordination Theory, to have a "human mind" is to be equipped with $(P, +)$. Such a characterization makes it possible to keep the objectivization of human cognition to a minimum: all humans start off from the same point with respect to the state of

---

combination of any two concepts will produce a concept, and we need not attribute different ontological status to "complex" and "simple" concepts/thoughts, and therefore need not define different *input* and *output* domains for the function modeling concept composition. Mathematically speaking, binary concept composition can thereby simply be defined as a closed operation under C, with domain: $C \times C \rightarrow C$.

[38] By approaching the mind as a computational system, Subjective Coordination Theory certainly does share some theoretical affinity with the Computational Theory of Mind (see Rescorla, 2020, for an overview) or the Language of Thought Hypothesis (LOT; Fodor, 1975, 2008). I say "some" here because I am not quite sure, for instance, of whether Subjective Coordination Theory would satisfy LOT's requirement of *systematicity*, given how strongly I am willing to commit to a contextualist and dynamic approach to cognition that would make it possible for the same linguistic expression to be associated with different conceptual representations across individuals, or across the same individuals at different points in time. And that is to say nothing of the principle of compositionality itself, and whether the framing of Axiom 2 would actually be considered acceptable by either theories in the first place. The question certainly deserves a thorough examination, but I leave that to future work.

[39] Nefdt (2020) distinguishes what he calls "process compositionality" from "state compositionality", which I believe captures a similar distinction:

> The most common kind of compositionality discussed in the literature [...] is what I will call *Process Compositionality*. The central idea is that the property of compositionality is located at the procedural level. What this means is that if a compositional procedure, such as a rule-to-rule mapping, is followed then the system in question is process compositional. [...] *State compositionality*, on the other hand, is a property of a structure identified by the possibility of decomposing that structure or state into smaller meaningful units. A helpful analogy is a puzzle here. A puzzler might have used particular heuristics to construct the overall picture (corners first, left to right, colour matching etc.) yet the state of the completed puzzle can be deconstructed (for later reconstruction, perhaps) in terms of other meaningful arrangements (ignoring the case of randomly deconstructing here). The state of a system itself can be said to compositional in this sense if it can be subdivided into meaningful parts. However, state compositionality is *theoretically* independent of process compositionality (although in many cases they do coincide) (Nefdt, 2020, pp.55–56).

their conceptual space (i.e. set of concepts) and every conceptual space should then evolve idiosyncratically as a function of experience—new concepts being added to the conceptual space as pre-existing ones are composed in response to experienced stimuli. Such considerations take inspiration from Chomsky (2000), notably when he tells us that:

> The only (virtually) “shared structure” among humans generally is the initial state of the language faculty. Beyond that we expect to find no more than approximations, as in the case of other natural objects that grow and develop.
>
> Chomsky (2000, p.30)

The Axiom of Underdetermination (Axiom 3) takes inspiration from the principle of linguistic underdetermination introduced earlier in Section 1.4.1, but extends it to all forms of behavior or stimulus at large—linguistic and non-linguistic alike. Axiom 3 applies both ways: it tells us that the same overt behavior could have been produced by many different mental processes (or correspond to many different mental representations, either across different individuals or within the same individuals across different points in time) and that any given external stimulus can also be processed mentally in many different ways. In short, it states that there exists a *one-to-many* mapping between behavior/stimulus and mind, across individuals and across time.

The Axiom of Making Sense (Axiom 4) and the Axiom of Least Effort (Axiom 5) provide the theoretical grounds for characterizing communication success in a non-normative subjectivism. In Subjective Coordination Theory, to interpret a communicated signal becomes the process of trying to make sense of that signal by doing one’s best using whatever one has at their disposal (conceptually). I take inspiration here from Wang (2019) who, upon reflecting on the definition of “artificial intelligence”, details what he takes to be the critical difference between minds and machines:

> My own opinion about the aim of AI started from the vague feeling that traditional computational systems are based on a design principle that makes them very different from the human mind, and that this principle can explain many other differences between the machine and the mind: A program is traditionally designed to do something in a predetermined *correct* way, while the mind is constructed to *do its best* using whatever it has. Consequently, absolute correctness or optimality of solutions should not be used as the design criteria for a mind-like system [. . .]
>
> Wang (2019, p.16)

Axiom 4 takes inspiration from the communicative principle of relevance (Definition 8) and the presumption of optimal relevance (Definition 9) although it generalizes them to all kind of external stimuli (linguistic and non-linguistic alike) and adopts a slightly different perspective, focusing on the *ability* of the *subject* rather than the *property* of the *stimulus*—following the considerations I developed in Section 1.3.3. Its underlying assumption is that the mind is constantly "in processing"—whether consciously or not—and that interpretation is a process that is triggered by the environment rather than mobilized consciously at will.

Axiom 5, finally, follows the longstanding Principle of Least Effort found in psychology and cognitive science at large—be it in (Zipf, 1949) or later on in (Sperber & Wilson, 1986/1995) with their relevance-guided comprehension heuristics (Definition 10)—and sets the fundamental constraint that guides the overall interpretation process. In Subjective Coordination Theory, to "make sense" of an external stimulus is to produce a subjective interpretation of that stimulus that minimizes as much as possible the cognitive cost of the interpretation process.

2.2.3 Operationalizing the axioms: vectors and projections

In Section 2.2.1, I mentioned that abandoning the presupposition of mutual understanding did not have to imply abandoning all corresponding theoretical devices (such as *similarity*) but only required that we rendered them optional rather than necessary. In this section, I would like to illustrate how, conversely, abandoning the presupposition of mutual understanding does no have to imply abandoning all corresponding *methodological devices* either. Focusing on the particular case of *vector* and *vector space*, I propose to illustrate how Subjective Coordination Theory can make use of such devices to operationalize its own axioms. In doing so, I intend to introduce some important clarifications that should facilitate the comprehension of the computational simulation of subjective coordination that I will introduce later on in Section 2.3.2.

The use of vector and vector space has a long history in the field of cognitive science, tracing back notably to the *connectionist* models of the 1990s attempting to model human cognition through of use of *neural networks* (see Buckner & Garson, 2019, for an overview). As Gastaldi (2021) explains, neural networks are, in a sense, nothing more than vector-processing devices taking vectors as inputs and outputting other vectors in turn:

> In their most elementary form, artificial [neural networks] can be seen as ways of transforming a vector (i.e., a list of numbers) into another vector, through successive parametrized transformations, [footnote omitted] each of which takes a vector as an input and yields a vector as an output that will be fed as the input vector of the next transformation. An input vector is thus fed into the network that will be transformed into another vector, which will be in turn transformed into a new one again and again until a final vector, which will be taken as the output of the procedure. Each one of those successive vectors is considered as a *layer*. The existence of multiple layers is what makes neural networks *deep*.
>
> Gastaldi (2021, pp.3–4)

The whole point of appealing to vector and vector space in cognitive science is that one can then mobilize the full power of linear algebra. Gärdenfors (2014), for instance, introduces his book on *The Geometry of Meaning* as follows:

> This book [...] focuses not on the geography of the mind but on its *geometry* [...] A central idea is that the meanings that we use in communication can be described as organized in abstract spatial structures that are expressed in terms of *dimensions*, *distances*, *regions*, and other geometric notions. In addition, I also use some notions from *vector algebra*.
>
> Gärdenfors (2014, p.36)

Historically, vector spaces have proven to be particularly useful methodological tools for psychological approaches to language and meaning, notably as they make it possible to model "similarity" as the *geometric distance* between vectors standing for concepts in the mind. Yet, as Definition 14 and Definition 15 make clear, *vector space* and *metric* are actually distinct mathematical objects which remain independent from one another. One can thus perfectly make use of vectors space as a methodological tool to model psychological concepts without necessarily having to specify any metric whatsoever, the same way that one can perfectly adopt a psychological approach to communication without necessarily having to rely on a theoretical concept of similarity.

Definition 14 (Vector space)
A *vector space* over a field[40] $\mathbb{F}$ is a set V with two operations: *addition*, mapping the Cartesian product[41] $V \times V$ to V, and denoted by $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x}+\mathbf{y}$; and *scalar multiplication*, mapping $\mathbb{F} \times V$ to V, and denoted by $(a, \mathbf{x}) \mapsto a\mathbf{x}$. Elements of the vector space V are called *vectors*, and the elements of the field $\mathbb{F}$ are called *scalars*. These operations must satisfy the following properties for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and all $a, b \in \mathbb{F}$:

(i) Commutativity of vector addition: $\mathbf{x}+\mathbf{y} = \mathbf{y}+\mathbf{x}$.

(ii) Associativity of vector addition: $(\mathbf{x}+\mathbf{y})+\mathbf{z} = \mathbf{x}+(\mathbf{y}+\mathbf{z})$.

(iii) Existence of an additive identity: There exists an element $\mathbf{0} \in V$ such that $\mathbf{0}+\mathbf{x} = \mathbf{x}$.

(iv) Existence of an additive inverse: For each $\mathbf{x} \in V$ there exists an element $-\mathbf{x} \in V$ such that $\mathbf{x}+(-\mathbf{x}) = \mathbf{0}$.

(v) First distributive law: $a(\mathbf{x}+\mathbf{y}) = a\mathbf{x}+a\mathbf{y}$.

(vi) Second distributive law: $(a+b)\mathbf{x} = a\mathbf{x}+b\mathbf{x}$.

---

[40] See (Humpherys et al., 2017, Definition B.2.1 p.659) for a formal definition of the concept of "field". Typical examples of fields include the set of real numbers $\mathbb{R}$ or the set of complex numbers $\mathbb{C}$.

[41] See (Humpherys et al., 2017, Definition A.1.10 (vii) p.630) for a formal definition of the concept of "cartesian product".

(vii) Multiplicative identity: $1\mathbf{x} = \mathbf{x}$.

(viii) Relation to ordinary multiplication: $(ab)\mathbf{x} = a(b\mathbf{x})$.

Adapted from Humpherys et al. (2017, Definition 1.1.3, p.4)

Definition 15 (Metric)
A *metric* on a set X is a map $d : X \times X \to \mathbb{R}$ that satisfies the following properties for all $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ in X:

(i) Positive definiteness: $d(\mathbf{x}, \mathbf{y}) \geq 0$, with $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.

(ii) Symmetry: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$.

(iii) Triangle Inequality: $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$.

The pair (X, d) is called a metric space.

Adapted from Humpherys et al. (2017, Definition 5.1.1, p.180)

In cognitive science and beyond, it is often enough to think of vectors as “arrows” with coordinates in the space or as “arrays” of (real) numbers. Gallistel (1990), for instance, clearly points at the coordinate/arrow-based characterization of vectors in his chapter on “Vector Spaces in the Nervous System” when he states that:

> The purpose of this chapter is to review neurophysiological data supporting the hypothesis that the nervous system does in fact quite generally employ vectors to represent properties of both proximal and distal stimuli. The values of these representational vectors are physically expressed by the locations of neural activity in anatomical spaces of whose dimensions correspond to descriptive dimensions of the stimulus. The term *vector space*, which refers to the space defined by a system of coordinates, has a surprisingly literal interpretation in the nervous system.
>
> Gallistel (1990, p.477)

Marcus (2001, p.194), on the other hand, defines a vector in his glossary as “An ordered array of numbers, such as $[0, 1, 7]$”, rejoining thereby the formalization used in computer science at large, where a vector is defined as a one-dimensional *array*, array itself defined as an ordered collection of elements of the same type (e.g. numbers, see Butterfield et al., 2016, pp.318 and 1140 for the definitions of “array” and “vector” in computer science).[42]

Yet, as the formal Definition 14 makes clear, a vector space is actually a much more complex mathematical object than what those standard characterizations suggest. In fact, returning to the formal definition of vector space proves particularly useful here, for it makes clear that a vector space is a powerful mathematical object that has some very

---

[42] Or as in the previous quote from (Gastaldi, 2021, pp.3–4).

interesting properties as far as Subjective Coordination Theory is concerned. First of all, a vector space is a set equipped with a binary law of composition (*addition*) which can serve to model the mechanism of compositionality between thoughts and concepts modeled as vectors, so as to satisfy the Axiom of Productivity (Axiom 1). Second, the very idea of defining vectors as "coordinates" in standard usage evokes the fact that a vector space is very often equipped with a *basis* (Definition 16).

> **Definition 16 (Basis)**
> A (linear) *basis* (or a *coordinate system*) in a vector space V is a set B of linearly independent vectors such that every vector in V is a linear combination of elements of B.
>
> Adapted from Halmos (1958/1987, p.10)

For example, in the typical case of distributional models of word meaning illustrated in Figure 2.1, word meanings are typically modeled as arrow vectors with coordinates in the *standard basis*: here $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$ and $e_3 = (0, 0, 1)$ in three dimensions. Modeling the mind as a vector space equipped with a basis thus allows each concept vector to be decomposable into a linear combination of basis/primitive vectors, satisfying thereby the requirement of the Axiom of Atomic Compositionality (Axiom 2).

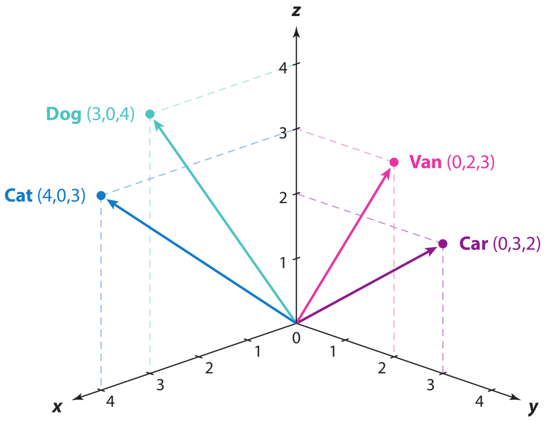


Figure 2.1: A three-dimensional representation of four word vectors in a semantic space, with their corresponding coordinates (borrowed from Lenci, 2018, Figure 1).

The methodological innovation of Subjective Coordination Theory, then, arises from appending to this traditional use of vector space the notion of (mathematical) *projection* (Definition 17) so as to operationalize the Axiom of Underdetermination (Axiom 3). The idea being that the mapping between messages/concepts/interpretations and signals is best formalized as a mathematical projection, that is, as a mapping from a space of "higher" dimensionality to a space of "lower" dimensionality, so that information is crucially "lost" by passing from the former modality to the latter. The underlying assumption is that signals are fundamentally "incomplete", in that their missing dimensions must be "enriched" to form full-fledged messages/concepts/interpretations. The interpretation process thus corresponds to a process of "pragmatic enrichment", where interpreters have to supplement missing information themselves through subjective interpretation.

> **Definition 17 (Projection)**
> Let V be a vector space, S be a subspace of V and D its complemented subspace in V so that $V = S \oplus D$, meaning that:
>
> $$\forall x \in V \quad \exists!(y, z) \in S \times D \quad x = y + z$$
>
> The projection on S along D is the application:
>
> $$p : \left\{ \begin{array}{ccccccc} V & = & S & \oplus & D & \to & S \\ x & = & y & + & z & \mapsto & y \end{array} \right.$$

The easiest way to conceptualize this notion of projection is to imagine the case where a three-dimensional object is projected to a two-dimensional plan. Every "point" of the parallelepipeds in Figure 2.2, for instance, can be formalized as a vector with coordinates (x, y, z) which, once projected on the xy plane along the z axis, has coordinates $(x, y, 0)$. As Figure 2.2 crucially suggests, different vectors in an initial space can thus project to the exact same vector in a subspace: the three parallelepipeds in Figure 2.2, for instance, though distinct in either "size" or "location", all project to the exact same black parallelogram. Given how we "loose" information regarding the initial z coordinate of the vector upon projection, "reconstructing" a vector in the space from its projection in the subspace entails that there can exist more than one solution—an *infinite* number of solution in this case. Yet, as you can see, not anything goes: any possible solution is constrained by the direction of projection and must remain alongside the z axis here. Thus, projections allow us to straightforwardly incorporate some of the crucial considerations I developed in Section 1.5.3: they model *constraints* on interpretation that leave room for a potentially unbounded set of solutions. Communication, thereby, can become a matter of *constraint satisfaction*, and no longer need to be formalized as a *closed* problem, let alone a *puzzle*.

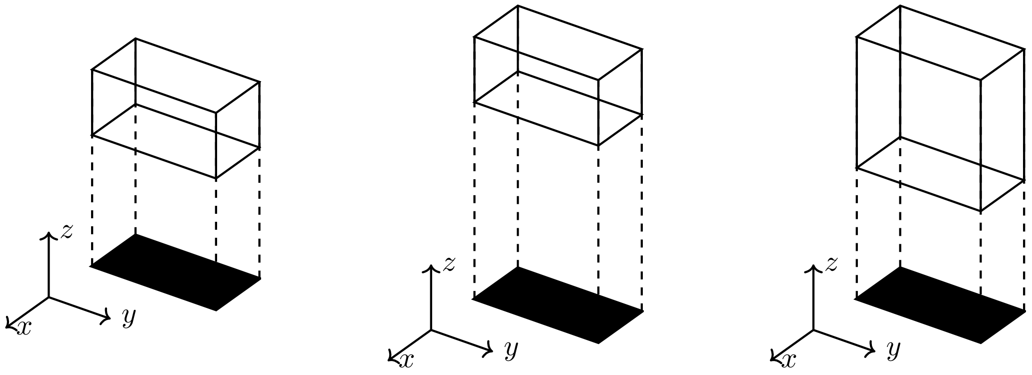


Figure 2.2: An example of three different parallelepipeds with different sets of coordinates, all corresponding to the exact same black parallelogram when projected on the xy plane along the z axis.

2.2.4 A theory of acommensurability in language and communication

Subjective Coordination Theory is a *psychological* theory of communication. As such, it focuses on how communication affects the *mind* of an individual, not their *behavior*. Its explanatory target is therefore communication *success*, not *effectiveness*. But it is also a theory that conceives communication success in a particular way: as the *correspondence* between interlocutors' minds rather than their *identity* or *similarity*—an idea which, I believe, can be found already in (Humboldt, 1836/1988) when he tells us that:

> Men do not understand one another [. . .] by mutually occasioning one another to produce exactly and completely the same concept; they do it by touching in one another the same link in the chain of their sensory ideas and internal conceptualizations, by striking the same note on their mental instrument, whereupon matching but not identical concepts are engendered in each.
>
> Humboldt (1836/1988, p.152)

This notion of correspondence, however, does not necessarily rule out identity or similarity: Subjective Coordination Theory is a theory of *a*commensurability in communication, not *in*commensurability. It is a theory that posits that meaning, language and interpretation *need not* be compared across interlocutors in order to characterize communication success, not that they *cannot* be compared. Its purpose is to change the epistemological status of some of the core assumptions that have grounded our theories of communication so far, not to argue that those assumptions are necessarily unfounded, or that they cannot be the object of proper scientific inquiry. Saying that "meaning", "language" or "interpretation" need not be the same across interlocutors to guarantee communication success does not preclude that they *could be* the same. But assessing whether they are the same requires first a theory of *objectivity*—a theory that can tell us how any two minds are to be compared in the first place, and what constitutes the independent and objective criterion we should rely on to do so. It requires then a theory of *context*—a theory that can tell us when, where and how the differences between those minds actually matter, and when they do not and can be safely turned into similarities, or identities. And it requires last a theory of *experience*—a theory that can tell us why and how those minds are to be found in the same configuration at a given point in time.

In the absence of any such theories, then, Subjective Coordination Theory applies a preventive *methodological individualism* which has it consider *a priori* that any experience grounding the content of a given mind is necessarily singular, and dynamic. This methodological individualism, however, does not imply the disconnection between subjectivity and experience, social or otherwise: in Subjective Coordination Theory, interpretation is always but the particular encounter between a given stimulus and a particular background of experiences. Neither does it imply that subjectivity entails full consciousness of one's mind, of its state, functioning, or content: Subjective Coordination Theory is ultimately a theory of subjectivity without *ego*.

Now, whether or not any two interlocutors actually end up interpreting "the same thing" in communication, the point is: not anything goes. Interpretation is a constrained process, and communication is a constraint satisfaction problem between interlocutors sharing those constraints. "Shared constraints", however, does not necessarily entail "shared language". First of all, because the constraints on interpretation are never reducible to language itself: the linguistic signal is always only part of the story. Second, because language acquisition is in itself pretty much of an open problem as well. Constrained as it may be by the environment of a speaker/signer (including their *social* environment of course), a language always finds itself in a singular configuration in the particular mind of an individual. Language, then, is but a matter of *idiolects*, and building a case for "shared language" from an idiolectal perspective would once again require an objective theory of shared experience which Subjective Coordination Theory does not provide. Be that as it may, the idiolectal approach to language adopted by Subjective Coordination Theory remains primarily rooted in its methodological individualism and its explanatory target: how *you* and *I* communicate with each other—not some hypothetical "average" speaker/signer of our "public" language—irrespective of how "erroneous" or "partial" our idiolects might be with respect to that public language.

The very notion of "idiolects" here invariably evokes the opposition with *conventional* approaches to language and meaning (Barber & Garcia Ramirez, 2021) reminiscent itself of a debate tracing back at least to Plato's *Cratylus* between the "conventional" and the "natural" in language and meaning.[43] In this traditional opposition, the "conventional" refers to what is *arbitrary* and could have been *otherwise* since it remains a matter of *choice* and, more importantly, of *collective* choice, even if only *tacit*. In contrast, the "natural" refers to what is *determined*, but more importantly to what *we* cannot *change*, and that in any case will not change any time soon: in other words, to what remains the same for everyone, every time, and to what we have no control over.

But conventions too can be the realm of "stability" against "change". In Lewis' (1969/2002) characterization, at least, conventions are behavioral regularities that constitute solutions to coordination problems and, as such, should necessarily *persist* in the absence of external interferences, given how deviation from the established convention is invariably costly for everyone. Lewis' theory is a theory of *equilibrium* and, as such, a theory of *stability* that fundamentally relegates the idea of change to the *accidental* and the *contingent*. Biology,

---

[43] In this debate, Hermogenes defends a conventionalist approach to word meaning where the association between a word and its referent is arbitrary and driven by *usage*:

> [N]o one is able to persuade me that the correctness of names is determined by anything besides convention and agreement. [...] No name belongs to a particular thing by nature, but only because of the rules and usages of those who establish the usage and call it by that name. (Plato, 1997, 384c–384d, p.103)

while Cratylus defends a naturalist approach to word meaning where the association between a word and its referent is driven by the *essence* of the thing the word denotes:

> A thing's name isn't whatever people agree to call it—some bit of their native language that applies to it—but there is a natural correctness of names, which is the same for everyone, Greek or foreigner. (Plato, 1997, 383a–383b, p.102)

on the other hand, can also be the realm of *life*, of permanent, continuous and intrinsic *change*, not only across *species*, but also across *individual organisms* themselves. Let us remember here indeed what Darwin (1859/2008) originally stressed about his approach to evolution through natural selection:

> No one supposes that all the individuals of the same species are cast in the very same mould. These individual differences are highly important for us, as they afford materials for natural selection to accumulate, in the same manner as man can accumulate in any given direction individual differences in his domesticated productions. [. . .] Hence I look at individual differences, though of small interest to the systematist, as of high importance for us, as being the first step towards such slight varieties as are barely thought worth recording in works on natural history. And I look at varieties which are in any degree more distinct and permanent, as steps leading to more strongly marked and more permanent varieties; and at these latter, as leading to sub-species, and to species.
>
> Darwin (1859/2008, pp.37;42)

Subjective Coordination Theory thus invites us to move past the traditional dichotomy between the *biological* and the *social*—or between the *natural* and the *cultural*—and to embrace the foundational opposition between the *static* and the *dynamic*. More than a theory of communication, then, Subjective Coordination Theory is a theory of the *dynamics* of communication: a theory that tells us that meaning, language and interpretation are dynamic entities, constantly changing as functions of occasional context and idiosyncratic experience, always potentially distinct from one point in time to the next, and from one individual to the next. But more fundamentally, it is a theory that intends to shift our approach to subjectivity in communication, and to no longer treat it as an *obstacle* on our way to success, or as a *problem* we need to solve. As such, it is a theory that opposes classical views which tell us that:

> We need a philosophy of mutual understanding, protecting shared understanding in the face of divergent ways and experiences.
>
> Blackburn (1984, p.8)

And as such, it must anticipate the inevitable critique of those classical views, which will invariably argue that questioning mutual understanding is necessarily pointless, and most certainly not a serious and productive scientific endeavor, for:

> After all, if we cannot in fact understand what others say or write and if they cannot understand us, it seems natural to conclude that each of us is little more than a psychological island: that is, we are isolated solipsists who hear only the echo of our own voices, all the while believing and acting under the tragicomic illusion that we are hearing and being heard by others. With such a conclusion as the only apparent alternative, it is not surprising that language

> theory has consigned the discussion of sceptical doubts about communicational understanding to the realm of non-serious discourse.
>
> Taylor (1992, p.3)

The best way to address those critiques, I believe, is to demonstrate concretely how Subjective Coordination Theory can prove scientifically productive, so let me now turn to this question and show you how Subjective Coordination Theory can account for some of the the fundamental phenomena we wish to explain when theorizing about communication.

2.3 Subjective coordination in action

2.3.1 Intuitions of communication success

As I mentioned already in Section 1.5.1 when introducing the *mutual knowledge hypothesis*, normative theories of communication that acknowledge the structural underdetermination of linguistic signals and the potential pervasiveness of subjectivity have no other choice but to posit the existence of specific cognitive devices that enable interlocutors to bridge the gap between their mind and that of others if they wish to accommodate the presupposition of mutual understanding. How could interpreters otherwise manage to retrieve the "correct" interpretation of an underdetermined linguistic signal? Mutual knowledge—and behind it the ability to represent the *beliefs* as much as the *knowledge* of others—is actually but *one* of the set of cognitive abilities that purportedly enable humans to bridge that mental gap between self and others, all part of what I call the "principle of intersubjectivity" (Definition 18).

Definition 18 (Principle of intersubjectivity)
Human beings are equipped with a set of cognitive abilities that enable them to successfully represent others' mental state and content. Those include:

1. The ability to share and represent others' *emotions* (i.e. *empathy*);
2. The ability to share and represent others' *desires*;
3. The ability to share and represent others' *attention* (i.e. *shared* or *joint attention*);
4. The ability to share and represent others' *goals* and *intentions* (i.e. *joint intention*);
5. The ability to share and represent others' *beliefs* and *knowledge* (i.e. *Theory of Mind* and *common knowledge*);

Adapted from Gärdenfors (2014, especially §3.1)

On the other hand, and as we saw in Section 1.5.2, if interlocutors are supposedly able to bridge that mental gap between self and others, it is also because it is presumed to be rather *small* in the first place. As Barsalou (2017) tells us indeed:

> [D]ifferent individuals have similar bodies, brains, and cognitive systems; they live in similar physical environments; they operate in highly-coordinated social contexts. As a result, different individuals acquire similar distributed networks for a given concept over the course of development. Within a particular social group or culture, different individuals' networks are likely to be highly similar, given similar coordinated experiences with many shared exemplars. Even across different cultures, these networks are likely to be highly similar, given that all humans have similar bodies, brains, and cognitive systems, operating in similar physical and social environments.
>
> Barsalou (2017, p.15)

The assumption of *shared meaning* thus takes place within a broader principle which I call the "principle of (conceptual) homogeneity" that posit *a priori* "sameness" between interlocutors' minds (Definition 19).

> **Definition 19 (Principle of homogeneity)**
> Interlocutors should share significant conceptual background given that they all live in the same world, have similar bodies, similar brains and similar cognitive systems, and should share similar social experiences grounding their conceptual representations.

Both principles constitute, in a sense, the two theoretical devices that serve to marginalize subjectivity within the mutual understanding paradigm. The principle of homogeneity marginalizes subjectivity *quantitatively* by saying that it should not be that significant in the first place; while the principle of intersubjectivity marginalizes subjectivity *qualitatively* by saying that, even if it ultimately proves to be quantitatively significant, it should not really matter as it should not prevent us from understanding one another anyway. Those two principles actually work hand-in-hand and culminate into what I call the "virtuous circle of mutual understanding" (Definition 20). The general idea being that shared meaning enables mutual understanding, which reinforces it in turn. Linell (2017) makes that point clear when he tells us that:

> Many scholars have taken intersubjectivity as a necessary assumption or a definitional point-of-departure for successful communication. [...] In addition, interlocutors must surely enter the conversation with something in common with regard to the topics raised, and their interaction will normally serve to increase the range of sharedness.
>
> Linell (2017, p.109)

> **Definition 20 (Virtuous circle of mutual understanding)**
> The more interlocutors successfully communicate with one another, the more conceptually homogeneous they should become; and the more conceptually homogeneous they are, the easier it should become for them to successfully communicate with one another.

The assumption is largely present in the cognitive science literature, especially within interactive approaches which formalize communication as a back-and-forth process of conceptual *alignment* between interlocutors (Pickering & Garrod, 2004, 2006, 2013, 2021). Garrod & Pickering (2009), for instance, tell us that:

> Alignment is typically achieved (to an extent that makes people believe that they generally understand each other), because people start off at a very good point. They communicate with other people who are largely similar to themselves, both because they process language in similar ways and because they share much relevant background knowledge.
>
> Garrod & Pickering (2009, p.294)

Meanwhile, Pickering & Garrod (2004, §2.4) tell us that "alignment at one level leads to alignment at another" so that we can expect successful alignment between interlocutors to increase conceptual homogeneity in turn, even if that process is expected to be slow and partial at best (Pickering & Garrod, 2006, §3.3).

Such theories—and more generally, all those which commit to the virtuous circle of mutual understanding—are what I call "theories of intersubjective coordination": they abide to the presupposition of mutual understanding and formalize communication as some form of *intersubjective* coordination accordingly, which manifests itself notably by their objectivizing communication success *as* mutual understanding. In the case of Pickering & Garrod (2004, 2006, 2013, 2021) mutual understanding becomes a matter of aligning "situation models" across interlocutors, which they define—following (Zwaan & Radvansky, 1998)—as multi-dimensional representations of the situation under discussion (encoding space, time, causality, intentionality and reference to main individuals under discussion) assumed to capture what people are "thinking about" during conversation (see Pickering & Garrod, 2004, §2.1, for details). The distinction they operate between *situation models*, *background knowledge* and *linguistic representations* allows them to easily accommodate the possible pervasiveness of subjectivity, all the while remaining within the paradigm of mutual understanding: communication can prove successful between interlocutors despite possible discrepancies across their respective background knowledge or linguistic representations, so far as their situation models can be interactively aligned during conversation (see Pickering & Garrod, 2006, §1). To them, then, the solution to underdetermination and subjectivity is to be found in the *interactive* nature of communication, and they consequently propose to revisit the standard characterization of communication conceived as a passive sender–receiver game between interlocutors—as typically found in the code model—and to replace it with a dynamic process of interactive alignment accordingly.

Subjective Coordination Theory, on the other hand, formalizes communication as *subjective coordination* plain and simple. Since it does not commit to the presupposition of mutual understanding, it need not objectivize communication success as the identicity (or even similarity) of interpretations, and need not operate a formal distinction between "situation models" and "background knowledge" or "linguistic representation"—or between "interpretation" and "meaning" more generally. In Subjective Coordination Theory, an *interpretation* is nothing more than a composition of pre-existing concepts (i.e. "meanings") and so a *concept* itself—and yet conflating the two notions does not prevent it from being able to account for the dynamic and contextual aspects of communication, as we will see throughout Section 2.3.2. More generally, abandoning the presupposition of mutual understanding means that Subjective Coordination Theory need not appeal to any such theoretical devices as the *principle of intersubjectivity* or the *principle of homogeneity* to account for communication success: in Subjective Coordination Theory, communication can be successful between interlocutors without them having to *successfully* represent each other's mind or be "conceptual homogeneous" in any way, since "success" is there defined in purely subjective terms and as the ability for an interpreter to find or construct an interpretation of the communicated signal at hand compatible with the potential singularity of their respective conceptual space.[44]

But how exactly is Subjective Coordination Theory then supposed to account for the intuition underlying the virtuous circle of mutual understanding? Changing the presuppositions of our theories is one thing, but we still need them to be able to account for the phenomena we would like to explain in the first place. Definition 20 may be framed here using some specific theoretical devices, it still intends to capture a fundamental intuition about communication, namely, that the more we communicate with each other, the more we seem to be able to understand one another.

The problem with that intuition, however, is that it involves the ability to straightforwardly evaluate mutual understanding which, as I discussed in Section 1.4.3, is far from given. First, because mutual understanding is not actually manifested behaviorally in communication, and because even if communicative effectiveness is, it offers no guarantee of mutual understanding itself as communicative effectiveness is neither necessary, nor sufficient to attest of mutual understanding. Second, because communication does not

---

[44] Which does not preclude that they *could* successfully represent each other's mind or that *some* form of conceptual homogeneity could actually exist between them. With respect to the principle of intersubjectivity, the point is not to negate the existence of a wide range of experimental evidence suggesting that human beings *can*, and often *do*, mobilize representations of each other's mental state and content during communication—be it during the production or the interpretation of utterances. The point here is only to stress that they need not do so *successfully*, which is the keyword in Definition 18 indeed. In Subjective Coordination Theory, interlocutors need not necessarily produce a representation of other minds that is "correct" in order for communication to proceed, and succeed (they can satisfy shared constraints on interpretation without necessarily converging to identical of even "similar" solutions/interpretations). This way, it should become possible to accommodate evidence of our "egocentric bias" in communication, and more generally accommodate the fact that *successful* intersubjectivity is far from given, and that communication often seems to proceed nonetheless, despite interlocutors lack of awareness of their own "failures" (see Nickerson, 1999; Keysar & Henly, 2002; Keysar et al., 1998, 2003; Kruger et al., 2005; Savitsky et al., 2011; Lau et al., 2022, and references therein for a comprehensive overview).

necessarily trigger overt and immediate behavioral response from interlocutors anyway, so that our intuitions of mutual understanding are more often than not grounded in the mere absence of overt conflict rather than the confirmed presence of agreement between us.

So, once again, we need to understand exactly what it is that we have the intuition *of* with those intuitions of mutual understanding. As I said in Section 1.5.3, we often seem to confuse intuitions of communicative effectiveness for intuitions of mutual understanding, but there are actually two separate claims in that statement. The first one is that we take our intuitions of communicative effectiveness to be intuitions of communication success. The second is that we take our intuitions of communication success to be intuitions of mutual understanding. My claim here is: we *do* have intuitions of communication success, but those intuitions are actually *not* intuitions of mutual understanding, and *not* intuitions of communicative effectiveness either. What are they intuitions of, then? Calling on to Quine (1990/1992) here can help, for he tells us that:

> Success in communication is judged by smoothness of conversation, by frequent predictability of verbal and nonverbal reactions [. . .] What is utterly factual is just the fluency of conversation [. . .]
>
> Quine (1990/1992, p.43)

What I propose to argue then is that our intuitions of communication success are, in fact, intuitions of the "smoothness" of our communicative interactions. That is, they are intuitions of the *cognitive cost* of those interactions; of the *processing effort* involved in interpreting our communicated signals; of how "easy" or "difficult" it is for each of us to make sense of one another. They are intuitions that pertain to the *process* of interpretation, not to its *output*. In short: they are intuitions of *understanding*, just not *mutual* understanding.

Just like mutual understanding, smoothness is clearly distinct from effectiveness: a communicative interaction can be smooth without it being effective, and vice-versa. Sometimes, our conversations can even have no other purpose than to be smooth. In such cases, there is no effectiveness in view, only success, which is where human communication fundamentally departs from signaling. Just like mutual understanding as well, smoothness corresponds to an intention to produce an effect *in the mind* of an interlocutor, not in their behavior: making communication smoother between us implies that we make it less cognitively costly for everyone. Unlike mutual understanding, however, smoothness *is* manifested behaviorally, and if the absence of overt and immediate behavioral feedback does not necessarily guarantee mutual understanding, it *does*, however, guarantee smoothness of communication: only the presence of overt and explicit behavioral conflict between us can call into question the smoothness of our interaction. Characterizing communication success in terms of processing cost provides a *psychological*, *gradual*, *non-normative* and most importantly *subjective* account of communication success, which still makes it possible to talk about "failure", but only in terms of (or due to) *cognitive limitations* from the interpreter. In Subjective Coordination Theory indeed, if communication fails between interlocutors, it is only be because the interpretation process for a particular communicated signal reaches a point beyond interpreters' cognitive capacities—remember the example of center-embeddings in (1.8)—and not because they could not retrieve the "correct" interpretation of that signal.[45]

By decoupling the question of communication success from that of mutual understanding, we open the door for alternative theories of communication that can better account for our intuitions of communication success. But more importantly, we get a better appreciation of the very phenomenon we are trying to explain. For indeed, behind the intuition of the virtuous circle of mutual understanding lies a much more straightforward intuition: that of a "virtuous circle of communication" plain and simple (Definition 21).

> **Definition 21 (Virtuous circle of communication)**
> The more interlocutors successfully communicate with one another, the easier it should become for them to successfully communicate with one another.

The virtuous circle of mutual understanding remains but a mere hypothesis after all. This virtuous circle of communication is the actual phenomenon that we need to account for.

2.3.2 Dynamics of communication: a computational simulation

My purpose in this section is to demonstrate how Subjective Coordination Theory can successfully account for the virtuous circle of communication (Definition 21) without having to mobilize any of the theoretical devices of intersubjective coordination theories such as the principle of intersubjectivity (Definition 18) or the principle of homogeneity (Definition 19). Doing so, I intend to show that the virtuous circle of communication need not be a virtuous circle of mutual understanding (Definition 20).

Given how accounting for the virtuous circle of communication requires making hypothesis about the *dynamics* of communication, I propose to explore the benefits of Subjective Coordination Theory through computational simulations, which prove particularly suited for the purpose at hand. I will first introduce my computational model in Section 2.3.2.1 so as to formalize the operational hypothesis in Section 2.3.2.2, then detail the experimental setup in Section 2.3.2.3 and report the results of my simulations in Section 2.3.2.4.

2.3.2.1 Computational model

I propose to formalize *communication* as an interactive process by means of which computational agents exchange (projected) conceptual representations (also referred to as "communicated signals" from now on for clarity) in an iterative fashion. I define an *agent* a_i by its concept matrix A_i, where a *concept matrix* A_i is characterized as a set of (row) *vectors* modeling the *concepts* present in the mind of the agent a_i. Following the considerations introduced in Section 2.2.3, I take every concept vector to be characterized as a linear combination of n primitive vectors $(p_1, \ldots, p_n)$, with $p_1 = (1, \ldots, 0), \ldots, p_n = (0, \ldots, 1)$.

---

[45] In the computational simulation of Section 2.3.2, I will consider for simplicity an idealized scenario where communication is always successful and just more or less costly.

Every concept matrix A_i will be initialized with the same set of primitives $(p_1, \dots, p_n)$. Equation (2.1) provides an example of a random concept matrix A, given $n = 3$ primitives.

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 3 & 1 \\ 3 & 5 & 9 \end{pmatrix} \tag{2.1}$$

As we can see, the matrix A in (2.1) is composed of the three primitives: $p_1 = (1, 0, 0)$; $p_2 = (0, 1, 0)$; and $p_3 = (0, 0, 1)$ and each row vector in A can be characterized as a linear combination of those primitives. For instance:

$$(1, 3, 1) = (1, 0, 0) + 3 \cdot (0, 1, 0) + (0, 0, 1) = p_1 + 3p_2 + p_3$$

I then define the *conceptual homogeneity* H between m agents as the Jaccard similarity between their respective concept matrices (Definition 22).[46]

Definition 22 (Conceptual homogeneity)
The conceptual homogeneity H of a set of m agents $a_1, \dots, a_m$ is the Jaccard similarity of their concept matrices $A_1, \dots, A_m$:

$$H(a_1, \dots, a_m) = \frac{|A_1 \cap \dots \cap A_m|}{|A_1 \cup \dots \cup A_m|} \tag{2.2}$$

Theoretically, H output values in $[0, 1]$:

- $H = 0$ means *null* homogeneity: *none* of the concepts are shared by *all* agents;[47]
- $H = 1$ means *full* homogeneity: *all* concepts are shared by *all* agents.

However, given that all agents will necessarily share the same set of primitives $(p_1, \dots, p_n)$, conceptual homogeneity can never be strictly null in practice: $H > 0$.

I then define *communicative cost* as follows (Definition 23).

Definition 23 (Communicative cost)
The cost of a communicative interaction is defined as the number of computation steps required by an interpretive agent to produce a concept matching the communicated signal.

---

[46] My main motivation for using the term *homogeneity* rather than *similarity* is to avoid confusion between a notion of conceptual similarity that applies to a whole *set* of concepts and one that applies to a *pair* of concepts. In practice, I use the term "conceptual homogeneity" to refer to the overlap between concept *matrices* and "conceptual similarity" to refer to similarity between two *single* concepts.

[47] Note that given the formalization of the Jaccard similarity, H can still be null even though one or more concepts are shared by two ore more agents in a larger population. For H to be non-null, *all* agents in the population must share at least one concept.

Definition 23 requires further explanations. In practice, communication is formalized as a computational simulation over a finite set of iterations. Each iteration models a particular communicative interaction between agents (see Figure 2.3 for an example with three agents). At each iteration, each agent randomly samples a single projected concept (i.e. communicated signal) from the matrix of another agent, and tries to make sense of it following the *making sense* algorithm (Algorithm 1). As Figure 2.3 indicates, several agents can potentially draw a different concept from the same agent at the same iteration, but each agent can only make sense of a single projected concept (i.e. communicated signal) at each iteration.

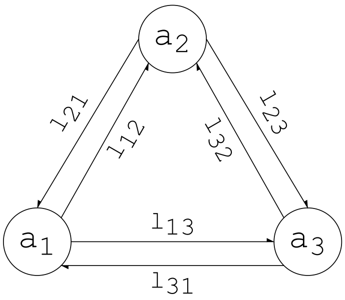


Figure 2.3: Communicative interaction between three agents (a_1, a_2, a_3). Arrows show all possible options for agents to sample a projected concept (i.e. communicated signal) from each other. Each agent can only sample a single concept per iteration: a_1 must choose between l_{21} and l_{31}, for instance.

Algorithm 1 Making sense

```
Input: $s$, $A$                              ▷ Input a signal $s$ and a concept matrix $A$
Output: $cost$                               ▷ Output the cost of making sense of $s$ given $A$
 1: function MAKE_SENSE_OF($s$, $A$)
 2:     $cost \leftarrow 0$                  ▷ Initialize cost value
 3:     $inter \leftarrow$ INTERPRET($s$, $A$)    ▷ Generate closest available interpretation
 4:     while PROJECT($inter$) $\neq s$ do   ▷ Until communication is successful
 5:         $inter \leftarrow$ REASON($inter$, $A$)   ▷ Apply reasoning to closest interpretation
 6:         $cost \leftarrow cost + 1$       ▷ Increment cost value by 1 each time reasoning is applied
 7:     end while                            ▷ End when communication is successful
 8:     MEMORIZE($inter$, $A$)               ▷ Memorize successful interpretation in $A$
 9:     return $cost$                        ▷ Return cost value of interaction
10: end function
```

Algorithm 1 specifies the overall interpretation process applied by agents to make sense of communicated signals, and how communicative cost is computed (line 6). Specifications for the INTERPRET, REASON and MEMORIZE functions in Algorithm 1 are detailed in the *interpretation* algorithm (Algorithm 2), the *reasoning* algorithm (Algorithm 3) and the *memorizing* algorithm (Algorithm 4). Concretely, the *making sense* algorithm (Algorithm 1) proceeds in three broad steps:

1. try and interpret the communicated signal s (line 3) by looking for the *closest* possible *interpretation* (see Algorithm 2 for details). The "interpretation" *inter* of a signal s is a concept which projection matches s: $\textsc{project}(inter) = s$ (see Algorithm 5 for details). "Closest" here refers to the minimal euclidian distance computed between vectors, as per the *distance* algorithm (Algorithm 6).
2. if the closest interpretation is not satisfactory (and until it is, see line 4) apply reasoning (Algorithm 3) to generate a new interpretation (line 5) and increment the communicative cost each time (line 6);
3. memorize the successful interpretation (Algorithm 4) by adding it to the agent's concept matrix A (line 8). Note that only the final successful interpretation is memorized, not the intermediate steps computed by the *reasoning* algorithm (see line 6 in Algorithm 3).

Algorithm 2 Interpretation

```
Input: $s$, $A$                                ▷ Input a signal $s$ and a concept matrix $A$
Output: $closest$                              ▷ Output the closest possible interpretation of $s$ in $A$
 1: function INTERPRET($s$, $A$)
 2:     $closest \leftarrow A[0]$                  ▷ Initialize closest concept as first concept in $A$
 3:     for all $c \in A$ do                       ▷ Iterate over concepts from oldest to newest
 4:         $proj \leftarrow$ PROJECT(c)
 5:         $dist \leftarrow$ DISTANCE($proj$,$s$)     ▷ The signal $s$ is already a projected concept
 6:         if $dist < mindist$ then
 7:             $mindst \leftarrow dist$
 8:             $closest \leftarrow c$
 9:         end if
10:     end for
11:     return $closest$                       ▷ Return the closest concept, not its projection
12: end function
```

Algorithm 2 and Algorithm 3 implement a simple brute-force search algorithm which attempts to find the best possible solution each time by iterating over all available concepts in the concept matrix.[48] This is an arbitrary modeling choice made so as to ease

---

[48] To guarantee deterministic search, order-sensitivity is added to set manipulation: all concept matrices are iterated over in reverse-order of added elements (from oldest to newest, see line 3 in Algorithm 2 and line 4 and line 5 in Algorithm 3) and every new element is stored at the "end" of the existing set (see line 2 in Algorithm 4).

Algorithm 3 Reasoning

```
Input: $inter$, $A$      ▷ Input an interpretation $inter$ and a concept matrix $A$
Output: $closest$        ▷ Output a new concept generated by combining two pre-existing concepts in $A$, so as to be as close as possible to $inter$
function REASON($inter$, $A$)
    $mindist \leftarrow \infty$            ▷ Initialize distance to highest possible value
    $ip \leftarrow$ PROJECT(inter)
    for all $x \in A$ do                   ▷ Iterate over concepts from oldest to newest
        for all $y \in A$ do
            $c \leftarrow x + y$           ▷ Generate a new concept by combining two pre-existing ones
            $cp \leftarrow$ PROJECT(c)
            $dist \leftarrow$ DISTANCE($ip$,$cp$)   ▷ Compare only projected concepts
            if $dist < mindist$ then
                $mindist \leftarrow dist$
                $closest \leftarrow c$
            end if
        end for
    end for
    return $closest$                       ▷ Return the new concept, not its projection
end function
```

Algorithm 4 Memorizing

```
Input: $c, A$            ▷ Input a concept $c$ and a concept matrix $A$
Output: $A$              ▷ Output the matrix $A$ updated with concept $c$
function MEMORIZE($c$, $A$)
    add $c$ to $A$       ▷ Apply order-sensitive set insertion: add $c$ to the end of $A$
    return $A$
end function
```

Algorithm 5 Projection

```
Input: $c = (c_1, \dots, c_n)$     ▷ Input a concept vector $c$ of dimension $n$
Output: projected $c$              ▷ Output its projected vector on $k$ dimensions, $k \leq n$
function PROJECT($c$)
    return $(c_1, \dots, c_k, 0, \dots, 0)$    ▷ $k$ is a hyparameter specified globally
end function
```

Algorithm 6 Distance

```
Input: $x = (x_1, \dots, x_n)$ $y = (y_1, \dots, y_n)$    ▷ Input two concept vectors $x$ and $y$ of dimension $n$
Output: $d(x, y)$    ▷ Output the euclidian distance between them
function DISTANCE($x$, $y$)
    return $\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
end function
```

the implementation process and that certainly does not aim at psychological plausibility. This decision, however, has no bearing on the considerations at hand: what matters to us here indeed is that all agents share the exact same search procedure (i.e. "cognitive functions"), not that this search procedure be the most efficient or the most cognitively plausible one. The same goes for the euclidian distance in Algorithm 6, which is chosen arbitrarily so as to formalize the success condition of the search function. Note that this distance function is an agent-internal mechanism which is *not* used as an objective criterion to quantify similarity across communicating agents' concepts. As such, it does not call into question the considerations introduced in Section 2.2.1. Following Subjective Coordination Theory, the computational model objectivizes the *constraints* on interpretation rather than interpretation itself. Consequently, it does not characterize communication success as the identicity or similarity between the interpreted and communicated concepts, but between their *projections*: "concept matching" in Definition 23 means identicity of *projected* concepts.

To better illustrate the overall functioning of the model, let me provide a concrete example. Consider communication occurring between two agents with concept matrices A (2.3) and B (2.4), which only differ by two concept vectors: A has $(2, 5, 2)$ which B does not have, and B has $(3, 8, 9)$ which A does not have. In this case, their homogeneity is $H = \frac{4}{6} = \frac{2}{3}$. Imagine now that B attempts to communicate $(3, 8, 9)$ to A (i.e., A samples $(3, 8, 9)$ from B) in a scenario where one dimension is projected. That is, B will send to A as communicated signal the projected vector $(3, 8, 0)$. A has no readily available interpretation to make sense of $(3, 8, 0)$: there is no concept c in A such that PROJECT$(c) = (3, 8, 0)$. A will thus proceed with *reasoning* by combining all possible pairs

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 3 & 1 \\ 2 & 5 & 2 \end{pmatrix} \tag{2.3}$$

$$B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 3 & 1 \\ 3 & 8 & 9 \end{pmatrix} \tag{2.4}$$

of vectors in A to find a suitable interpretation. As a matter a fact, it will be able to do so in a single step: combining $(1, 3, 1)$ and $(2, 5, 2)$ together will generate a concept vector $c = (3, 8, 3)$ such that $\textsc{project}(c) = (3, 8, 0)$. The communicative cost of this interaction will therefore be $cost = 1$. Note, however, that $(3, 8, 3) \neq (3, 8, 9)$: the communicated and interpreted concepts are not identical, and need not be so. Neither do they need to be intrinsically "similar" in any way. All they need to do is satisfy the same constraint: $\textsc{project}(c) = (3, 8, 0)$.

The purpose of my computational simulations, then, is to compare the dynamics of both *subjective* and *intersubjective* coordination. The way I formally distinguish them in the model is by considering that intersubjective coordination corresponds to the configuration where there is no projection: in practice, for every concept c, the *projection* algorithm (Algorithm 5) will just return the exact same concept vector c: $\textsc{project}(c) = c$.[49] My purpose in doing so is not to imply that intersubjective coordination does not incorporate any consideration over *underdetermination* whatsoever, but rather to stress that, under intersubjective coordination, interlocutors possess that means to overcome underdetermination so as to guarantee the identicity of their interpretations.[50] In doing so, I intend to emphasize that the difference between intersubjective and subjective coordination primarily lies in their characterization of success, and their objectivization of interpretation *versus* constraints on interpretation.[51]

Some may argue here that this constitutes an unfair characterization of intersubjective coordination: "situation models", for instance, would more likely resemble projected concepts given the formalization of (Pickering & Garrod, 2004), so that interactive alignment should actually correspond to my modeling of subjective rather than intersubjective coordination. Yet, as I will show later on in Section 2.3.2.4, only projection-free models of intersubjective coordination actually prove capable of accounting for the virtuous circle of mutual understanding, which motivates the present formalization.

2.3.2.2 Hypothesis

With those specifications in mind, we can now formalize a more operational hypothesis regarding the contribution of Subjective Coordination Theory to the characterization of the virtuous circle of communication. The original Definition 21 states that the more interlocutors successfully communicate with one another, the easier it should become for them to communicate successfully. Framed in computational terms, the virtuous circle of communication corresponds to the idea that the average communicative cost between

---

[49] This way, the difference between the intersubjective and subjective coordination scenario can be reduced, in terms of implementation, to a single hyperparameter value: the number of projected dimensions N_d (see Section 2.3.2.3).

[50] It might be useful here to stress that a communicated signal in the model does not necessarily correspond to what we would commonly treat as a single utterance: interpretation must be understood as the *output* of what could otherwise very well correspond to a back-and-forth process of communicative interactions between agents, and so to the exchange of *multiple* utterances.

[51] The success condition specified in line 4 of Algorithm 1 changes accordingly: in intersubjective coordination, communication is successful if and only if the interpreted concept is identical to the communicated one: $inter = s$.

agents should decrease as the number of iterations increases. In an idealized scenario where agents are able to communicate indefinitely with one another with no outside interference whatsoever, this implies that the average communicative cost should converge to zero after a sufficient number of iteration:

$$\lim_{iter \to \infty} \sum_{a_1}^{a_m} cost(a_i, iter) = 0 \tag{2.5}$$

Equation (2.5) is what I otherwise refer to as "persistent costless communication".

The "intersubjective coordination hypothesis" (Definition 24) then states that the virtuous circle of communication is necessarily a virtuous circle of mutual understanding.

Definition 24 (Intersubjective coordination hypothesis)
Full conceptual homogeneity between agents is necessary to guarantee persistent costless communication between them.

$$\lim_{iter \to \infty} \sum_{a_1}^{a_m} cost(a_i, iter) = 0 \implies H(a_1, \ldots, a_m) = 1 \tag{2.6}$$

On the other side, the "subjective coordination hypothesis" (Definition 25) states that the virtuous circle of communication need not be a virtuous circle of mutual understanding.

Definition 25 (Subjective coordination hypothesis)
Full conceptual homogeneity between agents is not necessary to guarantee persistent costless communication between them.

$$\lim_{iter \to \infty} \sum_{a_1}^{a_m} cost(a_i, iter) = 0 \centernot\implies H(a_1, \ldots, a_m) = 1 \tag{2.7}$$

My purpose in what follows is to validate the subjective coordination hypothesis, *contra* the intersubjective one, by showing that I can generate computational simulations of subjective coordination where the average communicative cost between agents converges to zero while conceptual homogeneity between them does not converge to one.

2.3.2.3 Experimental setup

Table 2.1 details the list of hyperparameters tested throughout the computational simulations, and their corresponding values. The term "initialization" refers to the initial iteration step before any communicative iteration has taken place ($iter = 0$). The parameter values are chosen as follows:

- $N_p = 10$, $N_c = 1000$ and $N_i = 10^4$ are chosen together so as to illustrate clearly the phenomenon of convergence of communicative cost to zero for $N_a = 2$, all the while minimizing the overall computation cost of the simulations so as to ease the replication process;[52]

---

[52] All simulations were designed so as to be able to run on a standard personal computer.

Parameter	Details	Values
N_p	Number of primitives	$N_p = 10$
N_c	Size of agents concept matrix at initialization	$N_c = 1000$
N_i	Number of iterations	$N_i \in \{10^4, 10^5\}$
N_a	Number of agents	$N_a \in \{2, 5, 10\}$
N_d	Number of projected dimensions	$N_p \in \{0, 1, 2, 5\}$
H_i	Homogeneity at initialization	$H_i \in \{0, .75, .85, .95\}$

Table 2.1: List of hyperparameters and their corresponding values.

- $N_i = 10^5$ is chosen so as to illustrate the phenomenon of convergence of communicative cost to zero for $N_a \in \{5, 10\}$, though note that most of the results will be presented for $N_a = 2$ so as to minimize the computation cost of the simulations;
- $N_d = 0$ corresponds to the intersubjective coordination scenario where no vector dimension is projected;
- $N_d \in \{1, 2, 5\}$ corresponds to the subjective coordination scenarios, and values are chosen so as to explore the influence of the number of projected dimensions on the experimental results presented;
- $H_i \in \{0, .75, .85, .95\}$ values are chosen so as to explore the influence of initial homogeneity between agents on the presented results. Note that $H_i = 0$ here is an approximation since homogeneity cannot be strictly null (hence my writing $H_i \sim 0$ in what follows).

To simplify the overall experimentation process, I set no upper boundaries to communicative cost. In practice, this means that communication is always successful, just more or less costly.

2.3.2.4 Results

The dynamics of communication is a matter of tendencies. A core feature of the computation model introduced in Section 2.3.2.1 is that it is highly contextual: the cost of any communicative interaction not only depends on the communicated signal, but also on the concept matrix of the interpreting agents, itself theoretically grounded in their past experiences. Concretely, what this means is that the average communicative cost between agents may vary significantly from one iteration to the next, and in rather unpredictable ways—as illustrated by the *without smoothing* curve in Figure 2.4.

Indeed, it is not because a group of agents have communicated smoothly at the previous iteration that they will necessarily communicated smoothly at the next. Successive costless interactions may very well be followed by a single costly interaction if the communicated

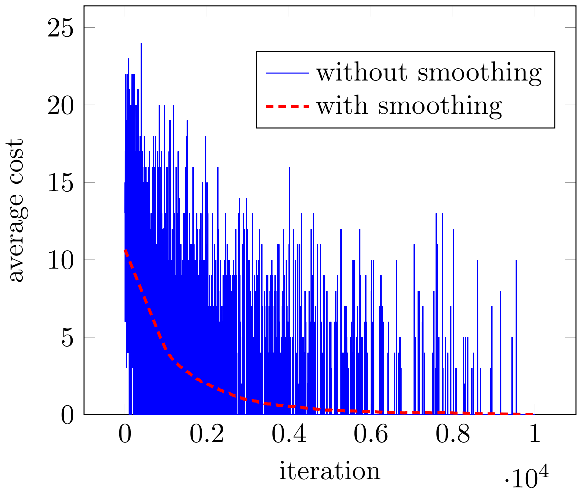


Figure 2.4: Average communicative cost across two agents, per iteration, *with* and *without* smoothing, in an intersubjective coordination scenario (no projection) with near-null initial homogeneity: $N_a = 2$, $N_d = 0$, $H_i \sim 0$.

signal and the interpreters' concept matrices are in such a way that the making sense algorithm requires a significant number of steps to successfully process that particular signal. Nevertheless, we can still observe general *tendencies* in the dynamics of communication, in terms of increase or decrease of average communicative cost across iterations, and in terms of convergence to particular values. However, such tendencies are somewhat obscured by the strong local variations of the raw data. For instance, the *without smoothing* curve in Figure 2.4 displays high average cost spikes after 6000 iterations, but those spikes often correspond to *single* costly iterations among hundreds of otherwise costless ones. There is actually a general tendency for the average cost to decrease and converge to zero which is not so clearly visible if you display only the raw *without smoothing* plot.

To better visualize those overall tendencies directly from the displayed plot, I apply *smoothing* to the raw experimental data in order to smoothen out those extreme local variations. I do so by relying on the Scipy (Virtanen et al., 2020) implementation of the Savitzky–Golay algorithm (Savitzky & Golay, 1964), with a window size of 1501 and a polynomial of order 1, chosen so as to obtain a continuous curve with limited local variations (see *without smoothing* curve in Figure 2.4). In what follows, I will only report smoothed plots rather than raw results so as to better compare the evolution of those tendencies across plots.

The virtuous circle of mutual understanding is constitent with intersubjective coordination. My first result is that the intersubjective coordination hypothesis holds, but only provided that we formalize communication as intersubjective coordination. That is, provided that communication success entails mutual understanding, persistent costless communication between agents necessarily entails full homogeneity between them.

Indeed, as Figure 2.5 shows, the *homogeneity* function converges to one when the *cost* function converges to zero. Communication does become more and more successful between agents as they communicate more and more with one another, and does become so as they become more and more homogeneous. Under intersubjective coordination, the virtuous circle of communication is necessarily a virtuous circle of mutual understanding.

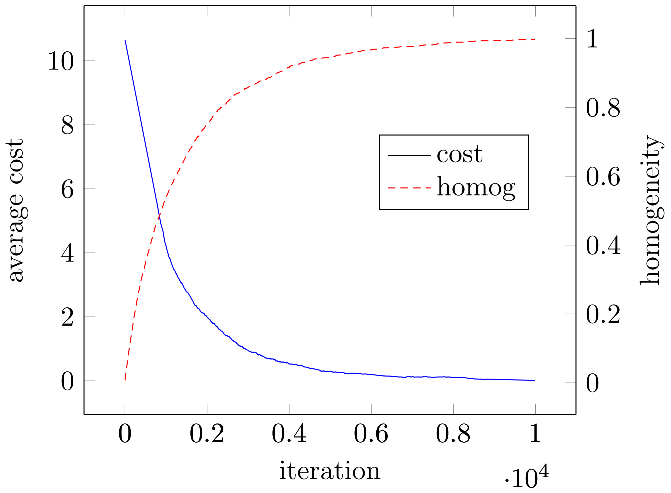


Figure 2.5: Average communicative cost across two agents (*cost*) and homogeneity (*homog*), per iteration, in an intersubjective coordination scenario (no projection) with near-null initial homogeneity: $N_a = 2$, $N_d = 0$, $H_i \sim 0$.

Those observations extend to simulations with more than two agents (see Figure 2.6 and Figure 2.7) although those require significantly more iterations to illustrate the same phenomenon. With ten agents, for instance, the simulation requires an increase of an order of magnitude in the number of iterations in order to reach persistent costless communication with full homogeneity ($N_i = 10^5$ instead of $N_i = 10^4$ for a two agents scenario).[53] Consequently, I only report simulations for a two agents scenario hereafter, but note that all results hold irrespective of the number of agents.

In the end, intersubjective coordination pushes the burden of proof over to characterization of those shared cognitive mechanisms that purportedly enable *successful* intersubjectivity between us. Provided that we *can* make sure that interpreting agents do recover the intended concept corresponding to the communicated signal, results show indeed that interlocutors will necessarily become more homogeneous as they communicate with one another. The intersubjective coordination hypothesis is at least coherent, although that does not make it necessarily true.

The virtuous circle of communication need not be a virtuous circle of mutual understanding. And indeed, further results suggest that it is *not* necessarily true. Figure 2.8 displays the core results of this section which validate the subjective coordination

---

[53] Note that the reduced rate of convergence of the homogeneity function in Figure 2.7 for five and ten agents is also an artifact of the Jaccard similarity (Equation 2.2) which requires that *all* agents share the same concepts to converge to one.

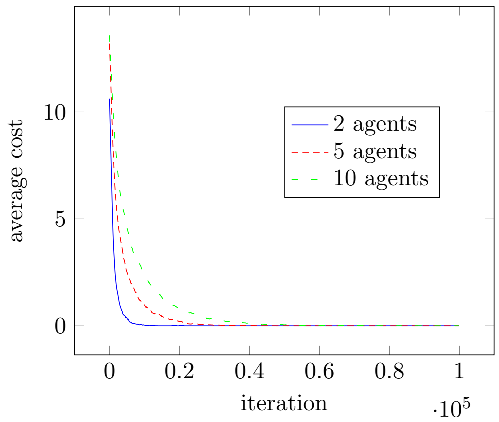


Figure 2.6: Average communicative cost across N_a agents, per iteration, in an intersubjective coordination scenario (no projection) with near-null initial homogeneity: $N_a \in [2, 5, 10]$, $N_d = 0$, $H_i \sim 0$.

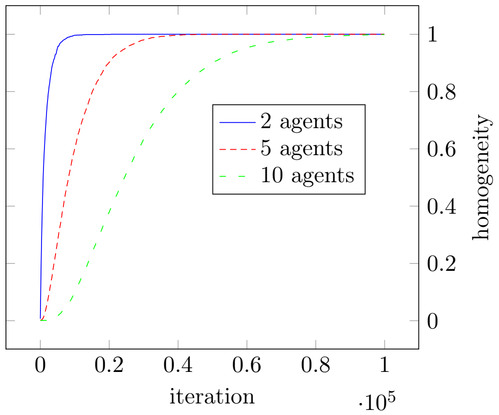


Figure 2.7: Homogeneity across N_a agents, per iteration, in an intersubjective coordination scenario (no projection) with near-null initial homogeneity: $N_a \in [2, 5, 10]$, $N_d = 0$, $H_i \sim 0$.

hypothesis. As it clearly illustrates, the virtuous circle of communication need not be a

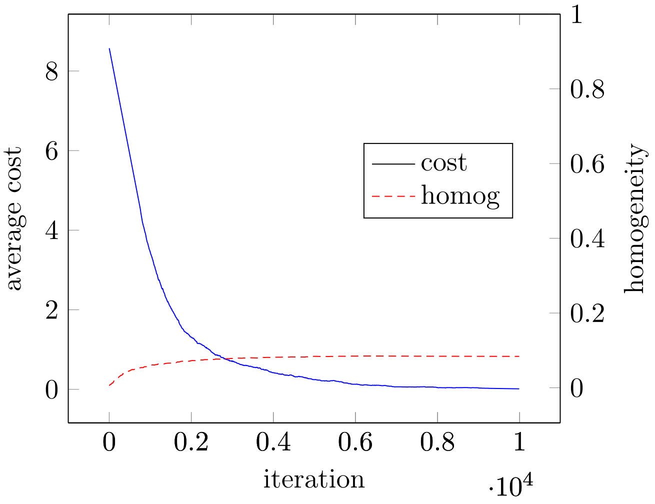


Figure 2.8: Average communicative cost across two agents (*cost*) and homogeneity (*homog*), per iteration, in a subjective coordination scenario with one projected dimension and near-null initial homogeneity: $N_a = 2$, $N_d = 1$, $H_i \sim 0$.

virtuous circle of mutual understanding: provided that we formalize communication as subjective coordination, persistent costless communication between agents proves possible despite the absence of full conceptual homogeneity between them. As Figure 2.8 illustrates, even though the average communicative cost between agents does converge to zero after a sufficient number of iterations, their homogeneity does not converge to one. Subjective coordination proves that full homogeneity is not necessary to guarantee persistent costless communication.

Those observations extend to various scenarios with different numbers of projected dimensions N_d. As Figure 2.9 and Figure 2.10 show, all scenarios lead to costless communication after a certain number of iterations, while only the intersubjective coordination scenario ($N_d = 0$) leads to full homogeneity. In all subjective coordination scenarios ($N_d > 0$), however, homogeneity need not be full to guarantee costless communication.

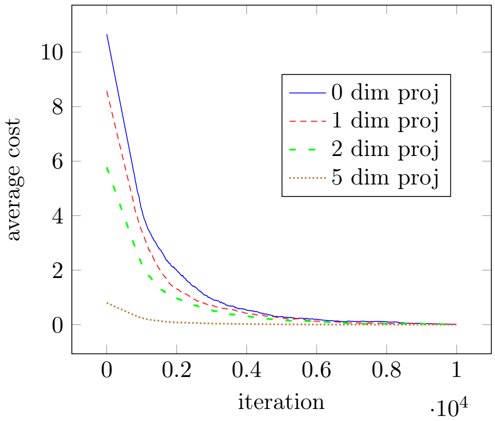


Figure 2.9: Average communicative cost across two agents, per iteration, in both intersubjective (no projection) and subjective coordination scenarios, with N_d projected dimensions and near-null initial homogeneity: $N_a = 2$, $N_d \in [0, 1, 2, 5]$, $H_i \sim 0$.

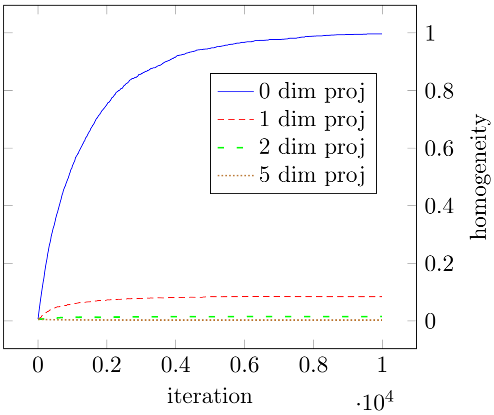


Figure 2.10: Homogeneity across two agents, per iteration, in both intersubjective (no projection) and subjective coordination scenarios, with N_d projected dimensions and near-null initial homogeneity: $N_a = 2$, $N_d \in [0, 1, 2, 5]$, $H_i \sim 0$.

In the end, what those results suggest is that calling into question the presupposition of mutual understanding does not have to imply giving up on our intuitions of communication success: subjective coordination proves perfectly capable of accounting for the virtuous circle of communication. But what they suggest more fundamentally is that subjectivity need not be seen as an obstacle to communication success, especially when success is framed in terms of smoothness of our communication interactions: it is perfectly possible, as we see here, to construct theories of communication that function *around* subjectivity—and not *despite* or *against* it—and still prove capable of capturing our fundamental intuitions about language and communication.

Homogeneity cannot compensate for underdetermination. Now, some may argue here that the results displayed in Figure 2.9 and Figure 2.10 are mere byproducts of initializing homogeneity at zero. After all, full conceptual homogeneity may not be *necessary* to guarantee persistent costless communication, but it should certainly prove *sufficient*, even in subjective coordination. If agents do share the exact same set of concepts, shouldn't they *de facto* reach mutual understanding in communication, even with projections/underdetermination? The hypothesis is attractive, but actually mistaken, or at the very least incomplete.

First of all, full homogeneity is not actually, in itself, a guarantee of mutual understanding in subjective coordination. Consider indeed the following example of two agents with concept matrices A (2.8) and B (2.9) which differ only in one respect: the order of their row vectors $(1, 3, 1)$ and $(1, 3, 7)$. Those two agents are in full homogeneity ($H = 1$)

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 3 & 1 \\ 1 & 3 & 7 \end{pmatrix} \tag{2.8}$$

$$B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 3 & 7 \\ 1 & 3 & 1 \end{pmatrix} \tag{2.9}$$

given that they share the exact same set of concept vectors. Yet, if B tries to communicate $(1, 3, 7)$ to A by projecting it to $(1, 3, 0)$ ($N_d = 1$ to keep things simple) A will interpret it as $(1, 3, 1)$ given that it will be the first available concept in its concept matrix to project to $(1, 3, 0)$. In other words: it is not because agents are perfectly homogeneous that they necessarily mutually understand each other.

Now, of course, the argument provided here is completely artifactual of the particular formalization of the computational model, and especially of the order-sensitivity of the search algorithm. Be the larger consideration still stands: for full homogeneity to guarantee mutual understanding, subjective coordination needs to carry out further hypotheses regarding the functioning of human cognition. In a certain sense, it needs a theory of human cognition that can guarantee interpretation to be insensitive to the *history* of agents' conceptual experiences. In the absence of any such theory, it remains impossible to confirm the original assumption.

Now, imagine that we *did* possess such a theory. Nobody would probably argue that full homogeneity constitutes a plausible scenario in communication: *some* form of idiosyncrasy across interlocutors' minds is always expected. So the real question, in fact, becomes that of whether a sufficiently high level of homogeneity can actually "compensate" for underdetermination so as to guarantee mutual understanding. The underlying intuition being that mutual understanding could still "happen" in subjective coordination—despite interlocutors having no proper means to guarantee the identicity of similarity of their interpretations—precisely because interlocutors are expected to be highly homogeneous in the first place.

Yet, once again, that hypothesis is not confirmed by the model. Consider this time the example of two agents with concept matrices A (2.10) and B (2.11) where all vectors in A are also in B, and all but one vector (the last) in B are also in A, in the exact same order. In a scenario where A has 19 concepts and B has 20, their homogeneity is $H = \frac{19}{20} = .95$. Imagine now that B tries to communicate $(3, 9, 5)$ to A by projecting it to $(3, 9, 0)$ and imagine that A has no readily available concept matching that projection. The reasoning algorithm will start composing concepts together until it reaches the combination $(1, 3, 1) + (2, 6, 3) = (3, 9, 4)$ which projection matches $(3, 9, 0)$, and yet, as you will

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 3 & 1 \\ 2 & 6 & 3 \\ \vdots & \vdots & \vdots \\ 2 & 5 & 2 \end{pmatrix} \tag{2.10}$$

$$B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 3 & 1 \\ 2 & 6 & 3 \\ \vdots & \vdots & \vdots \\ 2 & 5 & 2 \\ 3 & 9 & 5 \end{pmatrix} \tag{2.11}$$

note $(3, 9, 4) \neq (3, 9, 5)$. In other words: it is not enough for agents to share near-identical conceptual backgrounds and perfectly identical cognitive mechanisms for them to mutually understand one another: homogeneity cannot compensate for underdetermination in subjective coordination.

This particular example extends to computational simulations as well: all results in Figure 2.12 display a similar pattern of stagnating homogeneity despite systematic convergence to costless communication in Figure 2.11, irrespective of which value initial homogeneity is set to. Even in the scenario where initial homogeneity is the highest ($H_i = .95$) homogeneity does not actually converge to one when average communicative cost converges to zero, and actually decreases slightly to converge to $H = .93$.

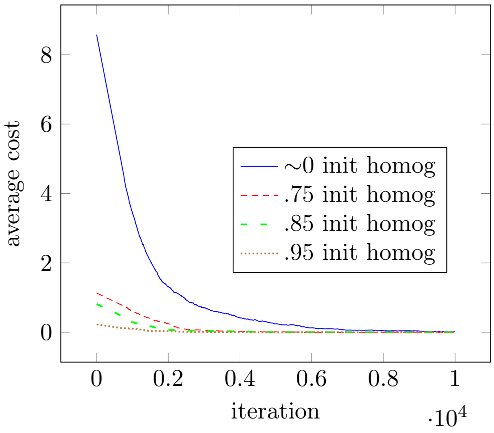


Figure 2.11: Average communicative cost across two agents, per iteration, in a subjective coordination scenario with one projected dimension and H_i initial homogeneity: $N_a = 2$, $N_d = 1$, $H_i \in [0, .75, .85, .95]$.

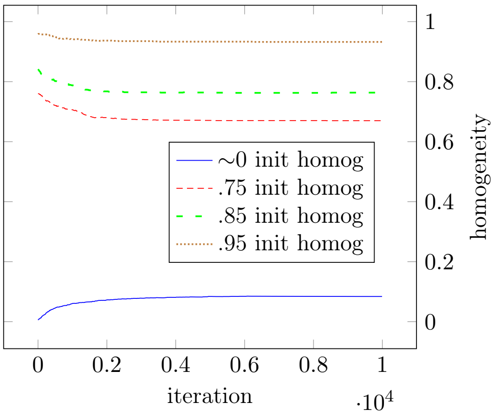


Figure 2.12: Homogeneity across two agents, per iteration, in a subjective coordination scenario with one projected dimension and H_i initial homogeneity: $N_a = 2$, $N_d = 1$, $H_i \in [0, .75, .85, .95]$.

In the end, all those considerations confirm that the principle of intersubjectivity takes precedence over the principle of homogeneity in accounting for the virtuous circle of mutual understanding. The principle of homogeneity *alone* cannot account for the virtuous circle of mutual understanding. There is no quick fix as far as subjective coordination is concerned: the virtuous circle of mutual understanding only holds if you first posit the existence of mutual understanding.

2.3.3 From subjective coordination to collective intelligence

The purpose of the previous Section 2.3.2 was to demonstrate the explanatory power of Subjective Coordination Theory; in effect, to show that it is concretely possible to abandon the presupposition of mutual understanding and still prove capable to account for some of our fundamental intuitions about communication success and its dynamics. The point of using computational simulations, however, was not necessarily to demonstrate what *is*, but rather to explore what *could be*; to help us get rid of unnecessary theoretical requirements and extend our range of possibilities in devising alternative theories of communication. Section 2.3.2, in particular, showed that neither the principle of intersubjectivity nor the principle of homogeneity were necessary to account for the virtuous circle of communication. But it also showed something more profound, namely, that underdetermination—broadly conceived—could not only serve to accommodate subjectivity, but also enable it to persist, if not flourish. Indeed, all the results of simulations experimenting with non-null initial homogeneity in Figure 2.12 show a slight decrease in homogeneity as agents reach persistent costless communication. In other words, successful subjective coordination actually *increases* agents' subjectivity.

It is one thing to argue that subjectivity is unavoidable in communication since interlocutors can never be expected to share the exact same set of background experiences, so that language must learn to make due with it if it is to function successfully and effectively as a communication system. But could there actually be good reasons for underdetermination to exist in the first place, other than to merely *accommodate* this state of fact? Could there be any reason indeed for underdetermination to enable subjectivity to persist, or even flourish, in that subjectivity *itself* would bring us benefits, *evolutionary* benefits, that would explain why language is structured the way it is—in a sense, as an *adaptation* to subjective coordination? And what would those evolutionary benefits of subjectivity be, then?

Tomasello (2014) tells us that humans fundamentally distinguish themselves from other great apes by a striking ability to adapt to their environment which has enabled them to migrate all over the globe when other great apes remain mostly located around the equator:

> Unlike other great apes, who all live in the general vicinity of the equator, modern humans have migrated all over the globe. They have done this not as individuals but as cultural groups; in none of their local habitats could a modern human individual survive for very long on his own. Instead, in each specific environment, modern human cultural groups have developed collectively a set of specialized and cognitively complex cultural practices to accommodate the

> local conditions, from seal hunting and igloo building to tuber gathering and bow-and-arrow making—not to mention science and mathematics.
>
> Tomasello (2014, p.120)

This remarkable adaptability—Tomasello (1999) tells us—derives from a species-specific ability for *cumulative culture*, that is, for a joint process of creative invention and faithful social transmission:

> The evidence that human beings do indeed have species-unique modes of cultural transmission is overwhelming. Most importantly, the cultural traditions and artifacts of human beings accumulate modifications over time in a way that those of other animal species do not—so-called cumulative cultural evolution. Basically none of the most complex human artifacts or social practices—including tool industries, symbolic communication, and social institutions—were invented once and for all at a single moment by any one individual or group of individuals. Rather, what happened was that some individual or group of individuals first invented a primitive version of the artifact or practice, and then some later user or users made a modification, an "improvement," that others then adopted perhaps without change for many generations, at which point some other individual or group of individuals made another modification, which was then learned and used by others, and so on over historical time in what has sometimes been dubbed "the ratchet effect" [(Tomasello et al., 1993)]. The process of cumulative cultural evolution requires not only creative invention but also, and just as importantly, faithful social transmission that can work as a ratchet to prevent slippage backward—so that the newly invented artifact or practice preserves its new and improved form at least somewhat faithfully until a further modification or improvement comes along.
>
> Tomasello (1999, pp.4–5)

The fundamental point being that humans are a uniquely cooperative species. Within this general paradigm, then, communication is taken to be a biological adaptation to cooperation and social interaction in general, given that:

> Cooperative communication [. . .] arose as a way of coordinating [. . .] collaborative activities more efficiently, first inheriting and then helping to build further a common psychological infrastructure of shared intentionality.
>
> Tomasello (2008, p.8)

"Shared intentionality" which Tomasello (2008) defines as:

> [. . .] what is necessary for engaging in uniquely human forms of collaborative activity in which a plural subject "we" is involved: joint goals, joint intentions, mutual knowledge, shared beliefs—all in the context of various cooperative motives.

> Tomasello (2008, pp.6–7)

Tomasello's considerations take as a starting point that no cultural product—be it material or symbolic—is never the sole product of a single human being. As such, it opposes the myth of the "lone genius" which has for long plagued the history of science and technology (Conner, 2005). It is a myth that runs deep indeed for, as Foucault tells us:

> [Michel Foucault:] The history of knowledge has tried for a long time to obey two claims. One is the claim of *attribution*: each discovery should not only be situated and dated, but should also be attributed to someone; it should have an inventor and someone responsible for it. General or collective phenomena on the other hand, those which by definition can't be "attributed," are normally devalued: they are still traditionally described through words like tradition, mentality, modes; and one lets them play the negative role of a brake in relation to the "originality" of the inventor. In brief, this has to do with the principle of the sovereignty of the subject applied to the history of knowledge. The other claim is that which no longer allows us to save the subject, but the truth: so that it won't be compromised by history, it is necessary not that the truth constitutes itself in history, but only that it reveals itself in it; hidden to men's eyes, provisionally inaccessible, sitting in the shadows, it will wait to be unveiled. The history of truth would be essentially its delay, its fall, or the disappearance of the obstacles which have impeded it until now from coming to light. The historical dimension of knowledge is always negative in relation to the truth. It isn't difficult to see how these two claims were adjusted, one to the other: the phenomena of collective order, the "common thought," the "prejudices" of the "myths" of a period, constituted the obstacles which the subject of knowledge had to surmount or to outlive in order to have access finally to the truth; he had to be in an "eccentric" position in order to "discover." At one level this seems to be invoking a certain "romanticism" about the history of science: the solitude of the man of truth, the originality which reopened itself onto the original through history and despite it. I think that, more fundamentally, it's a matter of superimposing the theory of knowledge and the subject of knowledge on the history of knowledge.
>
> Chomsky & Foucault (2006, p.15)

Behind the "ratchet effect" of Tomasello et al. (1993), then, we find the idea that the group is not necessarily the enemy of knowledge, or truth, and that it can also be the realm of *intelligence*, as the term "collective intelligence" suggests. It is an idea that is probably as old as Western philosophy itself, since we find it already in Aristotle and his *Politics* when he discusses *The Authority of the Multitude* in Book III, Chapter 11:

> [T]he view that the multitude rather than the few best people should be in authority would seem to be held, and while it involves a problem, it perhaps

> also involves some truth. For the many, who are not as individuals excellent men, nevertheless can, when they have come together, be better than the few best people, not individually but collectively, just as feasts to which many contribute are better than feasts provided at one person's expense. For being many, each of them can have some part of virtue and practical wisdom, and when they come together, the multitude is just like a single human being, with many feet, hands, 5 and senses, and so too for their character traits and wisdom. That is why the many are better judges of works of music and of the poets. For one of them judges one part, another another, and all of them the whole thing. [. . .] Taken individually, however, each of them is an imperfect judge.
>
> Aristotle (1998, 1281b)

Yet, a fundamental question remains: if the collective "helps" here—be it for Aristotle or Tomasello—is it solely by virtue of being of *collection* of individuals, or rather by virtue of being a collection of *different* individuals, that is, a collection of *subjectivities*? For indeed, as Graeber (2009) tells us:

> [I]f one is trying to solve a problem, or carry out a task, it's almost always going to be easier to do so with a group of five different people than with a group of identical clones.
>
> Graeber (2009, p.329)

So "collective intelligence", in fact, may be just another way to refer to the coordination of different subjectivities.

In cumulative culture, *creative invention* actually takes precedence over *faithful social transmission*: it is first because someone (or some group) comes up with an "innovation" that this innovation can be transmitted to others and further improved. So even if every cultural product is but a mere aggregate of tiny individual innovations—and never in its entirety the sole product of a single individual—the point is that it came to be the way it is because on many occasions, somehow, it got into the head of someone who conceptualized it in a certain way—in a *different* way—and because that conceptual difference proved to be relevant for the particular context where it took place. Tomasello (1999) focuses mostly on the mechanisms of social transmission, for he is primarily interested in characterizing what fundamentally distinguishes us from other primates, and tells us indeed that:

> Perhaps surprisingly, for many animal species it is not the creative component, but rather the stabilizing ratchet component, that is the difficult feat. Thus, many nonhuman primate individuals regularly produce intelligent behavioral innovations and novelties, but then their groupmates do not engage in the kinds of social learning that would enable, over time, the cultural ratchet to do its work [(Kummer & Goodall, 1985)].
>
> Tomasello (1999, p.5)

But there remains nonetheless an irreducible tension between creative invention and social transmission in the cumulative culture hypothesis: how are humans supposed to innovate collectively through successful cooperation if to do so they need to conceptualize the same cultural product differently in the first place?

Those who are ready to acknowledge the benefits of subjectivity for collective intelligence tell us that *some* form of objectivity is always required for humans to be able to act together. So is the case of Page (2007), for instance, who begins by introducing "toolbox diversity"—diversity is what he calls our "cognitive toolboxes", which include:

> Diverse Perspectives: ways of representing situations and problems
> Diverse Interpretations: ways of categorizing or partitioning perspectives
> Diverse Heuristics: ways of generating solutions to problems
> Diverse Predictive Models: ways of inferring cause and effect
>
> Page (2007, p.7)

Upon detailing the benefits of this toolbox diversity for collective intelligence, he tells us that:

> Up to this point, the results should bring joy and happiness. We might all think diversity is a wonderful thing. That's because we've ignored diverse preferences, differences in what we value. Preference diversity differs from toolbox diversity: Toolbox differences do not create conflict. Preference diversity can and does. For this reason, management books stress agreeing on a common goal—a common *fundamental* preference. If people disagree about what they're trying to accomplish, they function poorly as a collective.
>
> Page (2007, p.11)

And yet, as Graeber (2009) fundamentally argues:

> The fact that human beings live in incommensurable worlds has rarely prevented them from effectively pursuing common projects. It might seem contradictory—a philosopher might argue that if people live in incommensurable worlds, it is impossible for them to pursue the same ends because they could not even agree on what those ends even are—but this is the kind of objection that emerges from a world which starts from Platonic forms and tries to reason its way from those forms to explain empirical reality.
>
> Graeber (2009, p.329)

Graeber (2009) makes explicit what I have been touching upon implicitly throughout this thesis, namely, that the very "problems" we want our theories to solves are always tied to the metaphysical and ontological commitments of our systems of thoughts. In his ethnography of the *Direct Action Network*—a North American collective organized around the practice of consensus-based decision making—he explains that the practice of consensus itself does not just require changing our methods of collective organizing, but entails an entire paradigm shift with respect to what we take to bear metaphysical reality:

> [C]onsensus-based groups tend to avoid debating, let alone basing their identity, on questions of definition. Instead, they always try to bring things back to questions of action. So my first suggestion is that we look at this as if we were dealing with a political ontology that assumes that actions, and not objects, are the primary reality.
>
> Graeber (2009, p.328)

Under this new paradigm, then, subjectivity changes ontological status: it does not only become *unavoidable*, it also becomes *unproblematic*:

> [T]he practice of consensus decision-making [. . .] tends to make a veritable moral imperative of the need to integrate incommensurable perspectives [. . .] This is not to say that the average [. . .] meeting involves elaborate exercises in seeing things from other's points of view—in fact, the emphasis on shared projects of action makes it possible to largely side step such exercises, which are left for "trainings" and other educational events—but it presumes respect for incommensurable perspectives.
>
> Graeber (2009, p.520)

So, as we can see, subjectivity need not be seen as a problem for cooperation anymore that it needs to be seen as a problem for communication. It should therefore become possible to account for the intuitions underlying the "ratchet effect" of Tomasello et al. (1993) without necessarily having to mobilize any principle of "shared intentionality" whatsoever—just like it proved possible in this chapter to account for our intuitions of communication success without having to mobilize any principle of "intersubjectivity", or "homogeneity". Cooperation, just like communication, does not have to imply *intersubjective* coordination. The perspectives offered by Subjective Coordination Theory for evolutionary linguistics are therefore promising, given how it could simply dissolve (rather than resolve) some of the crucial problems posed by current theories, such as the critical tension underlying the cumulative culture hypothesis. There is ample room for scientific exploration beyond language and communication—though that may require, this time, interrogating the core presuppositions of our philosophy of action.

Back to our central question regarding the evolution of language: should we go as far as to consider that language evolved *for* the purpose of subjective coordination, given the evolutionary benefits purportedly conferred by subjectivity for collective intelligence? The hypothesis is plausible, and attractive. And yet, when theorizing about the evolution of human cognition, one must tread carefully. For as Lewontin (1998) reminds us indeed:

> We should not confuse plausible stories with demonstrated truth. There is no end to plausible storytelling.
>
> Lewontin (1998, p.129)

2.4 Language beyond communication

2.4.1 Function, evolution and adaptation

I began Chapter 1 with the debate over the communicative function of language, focusing notably on the role ambiguity could be said to play *for* or *against* viewing language as an efficient communication system. In my case, pondering on the communicative function of language lead me to focus on the proper characterization of the notion of "communication", but the original debate has actually focused on the separate though no less critical question of what constitutes the proper characterization of the notion of "function". It is on this particular question that I would like to close this Chapter 2 now: first, so as to complete the overall picture initiated in Chapter 1; and second, so as to illustrate concretely what kind of problems "plausible storytelling" could pose for the field of evolutionary linguistics at large.

Section 1.2.2 introduced the debate regarding the communicative function of language by mentioning the position of (Piantadosi et al., 2012) which argues, *contra* (Chomsky, 2002, 2008), that ambiguity actually makes language an *efficient* communication system. Piantadosi et al. (2012) build their case for the communicative efficiency of language by quoting the following extract from (Chomsky, 2002) at the beginning of their paper, which they take to illustrate the position they propose to argue against:

> The natural approach has always been: is [language] well designed for use, understood typically as use for communication? I think that's the wrong question. The use of language for communication might turn out to be a kind of epiphenomenon. I mean, the system developed however it did, we really don't know. And then we can ask: how do people use it? It might turn out that it is not optimal for some of the ways in which we want to use it. If you want to make sure that we never misunderstand one another, for that purpose language is not well designed, because you have such properties as ambiguity.
>
> Chomsky (2002, p.107)

Let me begin here by putting that quote in its original context, for I believe that it tells us much more than this extract would like to suggest. The excerpt is taken from an interview of Noam Chomsky by Adriana Belletti and Luigi Rizzi where they discuss the overall purpose of the Minimalist Program (Chomsky, 1995) and the specific question of the notion of optimization in language. Adriana Belletti and Luigi Rizzi begin by asking Noam Chomsky the following question:

> [Adriana Belletti and Luigi Rizzi:] The Minimalist Program explores the thesis that human language may be a "perfect system," a system optimally designed to meet certain conditions imposed by other cognitive systems that the language faculty interacts with. But what are the leading ideas about what would count as "perfection"? Some clarification is useful here. One can easily imagine

> criteria of perfection or optimality according to which human language would be far from optimally designed. Consider for instance the ubiquitous presence of ambiguity in natural language, a property which a "superengineer" would presumably avoid, given certain goals (to use a metaphor you often refer to in your minimalist writings). One could also argue that language, as an abstract computational capacity, is less than optimally adapted to the human performance system (with memory limitations, and so on), as it can give rise to all sorts of unusable structures (garden paths, center embedding, etc.), as you have often pointed out. Such criteria of optimal design are a priori conceivable and not unreasonable, but clearly they are not what is intended here. So, what kind of criteria of perfection make the minimalist thesis sustainable?
>
> Chomsky (2002, pp.105–106)

To which he replies:

> [Noam Chomsky:] Let's distinguish two questions. One is: what do we mean by optimality? Few rules is better than more rules, less memory used in computation is better than more memory used etc. There are some, not precise, general ideas about what optimality is. The second question is: what conditions is the system supposed to meet? I think what you're raising has to do with that question and you're absolutely right: there can be various points of view. If you take a standard functionalist point of view, you would ask: is the system designed for its use? So, is it going to be well designed for the uses to which people put it? And the answer there is "apparently not"; so the system does not seem to be all that well designed for use for the kind of reasons you mentioned (ambiguities, garden paths, lots of expressions that are unintelligible, expressions that are perfectly intelligible but not well formed). In some sense the system is not well designed for use, at least not perfectly designed for use, but it has to be designed well enough to get by. That's all that we discover: it's designed well enough to get by. That raises the question: can we find other conditions such that language is well designed, optimal for those conditions? I think we can, from a different perspective. So, instead of asking the standard functionalist question, is it well designed for use?, we ask another question: is it well designed for interaction with the systems that are internal to the mind? It's quite a different question, because maybe the whole architecture of the mind is not well designed for use. Let me see if I can make an analogy: take some other organ of the body, say, the liver. You may discover that the liver is badly designed for life in Italy because people drink too much wine and they get all sorts of diseases of the liver; therefore, the liver wasn't well designed for function. On the other hand, the liver might be beautifully designed for interaction with the circulatory system and the kidney and so on, and those are just different things. From the point of view of selection, natural selection,

> things must be well designed, at least moderately well designed for use, well designed enough so that organisms can reproduce and so on. But a totally separate question is: forgetting the use to which the object is put, is it well designed from the perspective of internal structure? That's a different kind of question, and actually a new one. The natural approach has always been: is it well designed for use, understood typically as use for communication? I think that's the wrong question. The use of language for communication might turn out to be a kind of epiphenomenon. I mean, the system developed however it did, we really don't know. And then we can ask: how do people use it? It might turn out that it is not optimal for some of the ways in which we want to use it. If you want to make sure that we never misunderstand one another, for that purpose language is not well designed, because you have such properties as ambiguity. If we want to have the property that the things that we usually would like to say come out short and simple, well, it probably doesn't have that property. A lot of the things we would like to say may be very hard to express, maybe even impossible to express. You often find that you can't express simple intentions and feelings that you would like to convey to somebody; a lot of personal interactions collapse because of things like that in ordinary life. So, the system is not well designed in many functional respects. But there's a totally separate question: is it well designed with regard to the internal systems with which it must interact? That's a different perspective and a new question; and that's the question that the Minimalist Program tries to answer.
>
> Chomsky (2002, pp.106–108)

What Chomsky (2002) is telling us here fundamentally is that we must distinguish *current utility* from *reasons of origin* when discussing the function of language, for the two may not be quite the same. And indeed, as Gould & Lewontin (1979) tell us in their seminal evolutionary biology paper on "The Spandrels of San Marco":

> One must not confuse the fact that a structure is used in some way [. . .] with the primary evolutionary reason for its existence and conformation.
>
> Gould & Lewontin (1979, p.587)

Which one of those, then, should we refer to when talking about the "function of language"? Under the *etiological view of functions*, talks about functions are talks about *origins* (see Sober, 1993/2000, p.85)—and for biological entities such as the language faculty, talks about functions are talks about *adaptations* (Definition 26).

Definition 26 (Adaptation)
Characteristic c is an adaptation for doing task t in a population if and only if members of the population now have c because, ancestrally, there was selection for having c and c conferred a fitness advantage because it performed task t.

Sober (1993/2000, p.85)

Millikan (1984/2001) herself refers to such evolutionary functions (or original/historical functions in the case of artifacts) as “proper functions”, so we can simply rephrase Chomsky’s (2002) argument using her terminology by saying that *communication may not be the proper function of language.*

Indeed, Chomsky (2002) takes the fact that language does not seem to be optimized for communication given ambiguity as proof that it could not have evolved *for* communication, and it is this assumption of a lack of communicative efficiency due to ambiguity that Piantadosi et al. (2012) propose to deconstruct. They do so notably by showing that communicative efficiency predicts that shorter words should be more ambiguous, given that they should prove easier to process, and confirm this prediction on English, German and Dutch. They interpret such evidence as proof of the existence of a pressure for communicative efficiency driving the structure of language, and more specifically the structure of the lexicon.[54] Yet, their study does not necessarily call into question Chomsky’s (2002) original argument. Indeed, evolutionary biology fundamentally distinguishes traits that are *adaptations* from traits that are merely *adaptive* (see Sober, 1993/2000, p.85). Sea turtles’ forelegs, for instance, can be *adaptive* for burying eggs in the sand without necessarily being *adaptations* for nest digging (Sober, 1993/2000, p.85)—and wings and feathers can prove adaptive for flight all the while being adaptations for thermoregulation (Longrich et al., 2012). The point being that a phenotypic trait may appear to be perfectly suited for the function it is currently being put to use without necessarily having been naturally selected for that function. Such kind of traits are what Gould & Lewontin (1979) refer to in their seminal paper as “spandrels”, and it is indeed on this critical consideration that Caplan et al. (2020) rely on to reply to Piantadosi et al. (2012) by arguing that:

> [T]he appearance of communicative efficiency with respect to lexical ambiguity is a spandrel.
>
> Caplan et al. (2020, p.2)

Language—efficient as it may be from the point of view of communication—may still not have evolved *for* communication.[55] Chomsky’s (2002) argument still stands: communication may not be the proper function of language.

---

[54] The idea in some sense traces back to Darwin (1871/2009) himself, who tells us that:

> A struggle for life is constantly going on amongst the words and grammatical forms in each language. The better, the shorter, the easier forms are constantly gaining the upper hand, and they owe their success to their own inherent virtue. (Muller, 1870, p.257) [as cited in (Darwin, 1871/2009, p.60)]

And explicitly connects it to natural selection:

> The survival or preservation of certain favoured words in the struggle for existence is natural selection. (Darwin, 1871/2009, pp.60–61)

[55] Caplan et al’s (2020) argument against functionalism could probably be reframed in more generic terms as the fact that *correlation is not (evolutionary) causality* although, in defense of functionalism, it is something it is well aware of. Levshina (2021, 2022), for instance, in her comprehensive review of

The considerations developed so far suggest that evolutionary linguistics could gain many useful insights from evolutionary biology—a point emphasized notably by Caplan et al. (2020) when appealing to (Gould & Lewontin, 1979). But the possible contribution of evolutionary biology does not stop at this distinction between adaptive traits and adaptations, or between current utility and reasons of origin. Gould & Lewontin (1979), in particular, provide a comprehensive critique of the whole adaptationist framework dominating evolutionary thinking:

> An adaptationist programme has dominated evolutionary thought in England and the United States during the past 40 years. It is based on faith in the power of natural selection as an optimizing agent. It proceeds by breaking an organism into unitary 'traits' and proposing an adaptive story for each considered separately. [. . .] We criticize this approach and attempt to reassert a competing notion (long popular in continental Europe) that organisms must be analysed as integrated wholes [. . .] We fault the adaptationist programme for its failure to distinguish current utility from reasons for origin [. . .] for its unwillingness to consider alternatives to adaptive stories; for its reliance upon plausibility alone as a criterion for accepting speculative tales; and for its failure to consider adequately such competing themes as [. . .] the separability of adaptation and selection [. . .] and current utility as an epiphenomenon of non-adaptive structures. We support Darwin's own pluralistic approach to identifying the agents of evolutionary change.
>
> Gould & Lewontin (1979, p.581)

---

functionalist work on the question of communicative efficiency, tells us explicitly that:

> [T]rade-offs, or inverse correlations, between linguistic variables related to different cues do not automatically imply efficiency as a driving force of language use and change, and the other way round. Levshina (2021, p.3)

Note, however, that even the fact that language could be adaptive for communication is not obvious. The fact that language change could be driven by language use, for instance (understood as use for communication) is notably challenged by the work of Labov (1972, 1994, 2001, 2010)—and that of Meillet (1921/1982) before him—who argues that:

> [N]o universal cause can account for the sporadic character of language change.
>
> Labov (2010, p.370)

And concludes more generally—*contra* (Darwin, 1871/2009, pp.60–61) in footnote 54—that:

> No linguist would today endorse this point of view, which runs exactly counter to our notion of the arbitrariness of the linguistic sign. Languages do not seem to be getting better and better, and we see no evidence for progress in linguistic evolution [(Greenberg, 1959/1974)]. Except for the development of vocabulary, we cannot argue for adaptive radiation in any area of language. The diversification of languages is not immediately and obviously functional, as the diversification of species may be. We receive no immediate benefit from not being able to understand the Russians or the Gaels, and the time taken to learn their languages does not seem to help in the survival of our own. We must seriously consider the possibility that the diversification of language is dysfunctional, and that we are worse off than if we all spoke a mutually intelligible version of post-Indoeuropean.
>
> Labov (1972, p.273)

Adaptationism, they tell us, is an entire *practice* of evolutionary thinking that entails a specific *attitude* towards the object of study—one which often translates as a form of "narrow teleology" which would like to see us consider each biological trait in isolation and treat it as an adaptation, that is, as being optimized to perform a specific function conferring a selective advantage to the organism as a whole:

> The adaptationist programme is truly Panglossian. Our world may not be good in an abstract sense, but it is the very best we could have. Each trait plays its part and must be as it is. [...] The admission of alternatives in principle does not imply their serious consideration in daily practice. We all say that not everything is adaptive; yet, faced with an organism, we tend to break it into parts and tell adaptive stories as if trade-offs among competing, well designed parts were the only constraint upon perfection for each trait.
>
> Gould & Lewontin (1979, pp.585–586)

So if Chomsky (2002) tells us that, from the point of view of current utility, language only needs to prove "good-enough" for communication, Gould & Lewontin (1979) also suggest more critically that we should refrain from the temptation of telling "just-so" stories about the evolution of language on the basis of apparent design if we do not want to fall into the "panglossian trap" of adaptationism yet again.[56]

Returning now to our discussion of Section 2.3.3 regarding the evolution of language and the evolutionary benefits of subjectivity, what such considerations tell us first is that language need not necessarily be approached as an adaptation to communication, and so to subjective coordination. But what they tell us also—and more fundamentally—is that the absence of a selective disadvantage does not necessarily have to imply the presence of a selective advantage. In sum, it is not because subjectivity is no longer seen as a *problem* for communication or coordination that it suddenly has to turn into a *benefit*. In the end, trying to build a case for underdetermination being an adaptation to subjective coordination given the selective advantage conferred by subjectivity for collective intelligence—attractive as the hypothesis may be—may prove to be nothing more than plausible storytelling. As far as the contribution of this chapter is concerned, saying that "subjectivity is not a problem" is probably "good-enough" already.

2.4.2 Towards a non-theological evolutionary linguistics

The field of evolutionary linguistics seems to have taken a particular turn over the past few decades, one which requires us to take side on what we consider to be the fundamental (political) instincts associated with human nature. The shared intentionality hypothesis of Tomasello (2008), for instance, rests on the assumption that humans are intrinsically *cooperative* and *altruistic* beings, and he sees this cooperative instinct as grounding both our communicative and social interactions in general:

---

[56] For a comprehensive version of this argument in the context of evolutionary biology, see also (Williams, 1966/1996, especially ch.1).

> [T]here must be some fairly specific connections between the fundamentally cooperative structure of human communication, as initially discovered by Grice, and the especially cooperative structure of human, as opposed to other primate, social interaction and culture in general.
>
> Tomasello (2008, p.xi)

It is actually this presupposition of an intrinsically cooperative or altruistic human nature that some of his critics, such as Reboul (2017), have called into question:

> [A] small minority of humans do seem capable of acts of great altruism, but this is hardly a dominant feature in humans. [. . .] This makes it highly unlikely that altruism could be the reason for human linguistic communication.
>
> Reboul (2017, p.181)

Those assumptions regarding our political nature as human beings are often grounded in the observation of hunter-gatherer societies, which are sometimes treated as possible mirrors of our past political selves, if not actually as direct evidence of our "true" human nature. Yet, as Reboul (2017) tells us:

> There is a persistent myth to the effect that hunter-gatherer societies are naturally peaceful and inter-or intra-group conflict in such societies is due to external intervention. [footnote omitted] This, however, seems to be false both for inter-and intra-group conflict in hunter-gatherer societies.
>
> Reboul (2017, p.179)

The debate seems to have found itself divided into two camps: one one side, those who follow Hobbes (1651/2010) in considering that our modern modes of social organizing have preserved us from our otherwise natural instincts for selfishness, deception, violence and domination; and on the other side, those who follow Rousseau (1754/1984) in considering that, on the contrary, it is those very modes of social organizing that have corrupted us away from our otherwise natural instincts for peace, kindness, cooperation and altruism. This opposition is what Boehm (1999) nicely summarizes as the polarization between "Hobbesian hawks" and "Rousseauian doves":

> As members of bands or tribes, humans can be quite egalitarian—particularly with respect to males. Yet we also develop degrees of despotism which, by mammalian standards, are truly staggering. This extreme range of behaviors can be bewildering. For a dwindling contingent of scholars, those who insist on unadulterated environmentalism, such disparities are taken to mean that our political nature is simply nonexistent. For many others, it remains a puzzle that must await the isolation of behavior genes in the laboratory. For still another contingent (including many of those mentioned above), it has been a subject of unresolved, philosophically oriented debate that seems to polarize the protagonists into Hobbesian hawks and Rousseauian doves.

> Boehm (1999, p.2)

Yet, as Boehm (1999) also suggests, there is a possible alternative to this binary framing of the problem: considering that there is no "political instinct" to naturalize in the first place. As a matter of fact, Graeber & Wengrow (2021), in their recent careful and exhaustive review of the anthropological and archeological evidence on the matter, tell us that:

> [I]t is clear now that human societies before the advent of farming were not confined to small, egalitarian bands. On the contrary, the world of hunter-gatherers as it existed before the coming of agriculture was one of bold social experiments, resembling a carnival parade of political forms, far more than it does the drab abstractions of evolutionary theory.
>
> Graeber & Wengrow (2021, p.4)

Pondering on the classical opposition between Hobbesian and Rousseauian, they tell us:

> It is basically a theological debate. Essentially the question is: are humans innately good or innately evil? But if you think about it, the question, framed in these terms, makes very little sense. 'Good' and 'evil' are purely human concepts. It would never occur to anyone to argue about whether a fish, or a tree, were good or evil, because 'good' and 'evil' are concepts humans made up in order to compare ourselves with one another. It follows that arguing about whether humans are fundamentally good or evil makes about as much sense as arguing about whether humans are fundamentally fat or thin.
>
> Graeber & Wengrow (2021, pp.1–2)

The question is, then: do such "theological" considerations really have their place in evolutionary linguistics? Does theorizing about the origins of human language and cognition necessarily requires grounding our hypotheses on fundamental assumptions about our (political) "human nature"? Could such assumptions really prove productive from a scientific perspective, given what Graeber & Wengrow (2021) fundamentally stress?

It is useful here, I believe, to turn to the famous debate between Noam Chomsky and Michel Foucault—debate incidentally titled "On Human Nature" (Chomsky & Foucault, 2006)—since Foucault actually clarifies what he takes to be the role of this notion of "human nature" in the overall scientific enterprise of the natural sciences, and its relation to theology in particular:

> [Foucault:] It is true that I mistrust the notion of human nature a little, and for the following reason: I believe that of the concepts or notions which a science can use, not all have the same degree of elaboration, and that in general they have neither the same function nor the same type of possible use in scientific discourse. Let's take the example of biology. You will find concepts with a classifying function, concepts with a differentiating function, and concepts

> with an analytical function [. . .] But there also exist "peripheral" notions, those by which scientific practice designates itself, differentiates itself in relation to other practices, delimits its domain of objects, and designates what it considers to be the totality of its future tasks. The notion of life played this role to some extent in biology during a certain period. [. . .] I would say that the notion of life is not a *scientific concept*; it has been an *epistemological indicator* of which the classifying, delimiting, and other functions had an effect on scientific discussions, and not on what they were talking about. [. . .] Well, it seems to me that the notion of human nature is of the same type. [. . .] **In the history of knowledge, the notion of human nature seems to me mainly to have played the role of an epistemological indicator to designate certain types of discourse in relation to or in opposition to theology or biology or history. I would find it difficult to see in this a scientific concept.**
>
> Chomsky & Foucault (2006, p.5, my emphasis in bold)

For Foucault, the notion of human nature—"peripheral" as it may be—is no less *structural* to the overall practice of science. In a certain sense, it is what sets the horizon of what our scientific enterprise is ultimately after. And in fact, it is his own theory of human nature grounded in the scientific concept of linguistic creativity that Noam Chomsky introduces first in that very same debate, prompting Foucault's response above. So the real question is not whether we *can* or *should* do without such a notion of "human nature" in evolutionary linguistics—all things considered, it seems that we cannot really do without it indeed—but rather whether we should construct it *in relation to* or *in opposition to*, theology, in our case. In practice, the question is whether we should include any "moral" consideration to our notion of human nature when theorizing about human evolution—as the classical opposition between Hobbesian and Rousseauian suggests.

It seems to me that we should be able to carry out research in evolutionary linguistics without having to take side here, that is, without having to posit the existence of an inherently "good" or "bad" human nature, and more generally without having to posit any form of rigid political or moral human nature whatsoever. Leaving theological and moral considerations aside is for me as much a matter of *good* science than it is of *ethical* science, since as Foucault tells us later on in that same debate:

> [Foucault:] [. . .] you can't prevent me from believing that these notions of human nature, of justice, of the realization of the essence of human beings, are all notions and concepts which have been formed within our civilization, within our type of knowledge and our form of philosophy, and that as a result form part of our class system [. . .]
>
> Chomsky & Foucault (2006, pp.57–58)

A statement which we should probably put into the more general context of his overall philosophical enterprise:

> Since its beginnings with Socrates, philosophy has typically involved the project of questioning the accepted knowledge of the day. Later, Locke, Hume, and especially, Kant developed a distinctively modern idea of philosophy as the critique of knowledge. Kant's great epistemological innovation was to maintain that the same critique that revealed the limits of our knowing powers could also reveal necessary conditions for their exercise. What might have seemed just contingent features of human cognition (for example, the spatial and temporal character of its perceptual objects) turn out to be necessary truths. Foucault, however, suggests the need to invert this Kantian move. Rather than asking what, in the apparently contingent, is actually necessary, he suggests asking what, in the apparently necessary, might be contingent. The focus of his questioning is the modern human sciences (biological, psychological, social). These purport to offer universal scientific truths about human nature that are, in fact, often mere expressions of ethical and political commitments of a particular society. Foucault's critical philosophy undermines such claims by exhibiting how they are the outcome of contingent historical forces, not scientifically grounded truths. Each of his major books is a critique of historical reason.
>
> Gutting & Oksala (2022, p.5)

One can only foresee the potentially dreadful consequences of naturalizing otherwise narrowly culturally and historically situated social habits, by granting them the virtue of being biological adaptations, purportedly "universal" and "natural". Theorizing about human nature is theorizing about what is necessarily shared and common to us all as human beings, so we should remain careful not to claim universality for our socio-historical and socio-cultural particularities. The frontier is always thin between *bad* scientific practice and *dangerous* pseudo-science, so that adopting a non-theological linguistics seems to me here more than just an epistemological necessity grounded in the will to guarantee the explanatory power of our scientific theories—it is, first and foremost, an *ethical necessity.*

2.5 Intermediate conclusion

I began Chapter 1 with the question of the communicative function of language, contributing to the debate first by proposing to distinguish ambiguity from ambiguous *system* (Section 1.3.1). There, I rejoined the traditional functionalist argument against the separation between (language) *structure* and *use*, and argued more specifically that it was this arbitrary separation and its underlying epistemological assumptions that actually generated the problem of ambiguity *itself*—so that ambiguity could be considered, in the end, a purely *artificial* problem from the point of view of communication (Section 1.3.2).

Do such considerations make me a functionalist, then? Not exactly. First of all, because, as we have seen, functionalist arguments about the communicative efficiency of language presuppose communicative effectiveness and through it mutual understanding altogether—which is precisely the assumptions I wanted to question in those first two chapters. To the functionalist argument that language is optimized for communication I thus oppose the following question: *for which communication*? The *intersubjective* and thus *objective* coordination between minds, or the *subjective* coordination of minds plain and simple? Second, because functionalist arguments about the communicative efficiency of language take place within a broader *adaptationist* paradigm which, as we have seen, deserves some questioning as well.[57] Ultimately, and as Chomsky (2002, pp.106–108) himself suggests, the question may not be so much *what* constitutes the function of language—and whether that is *communication* or something else (e.g. "internal thought", as per Reboul, 2017)—but whether language could really be said to have any function *at all*.[58] The question I started this thesis with—what is the function of language?—may thus be but a false problem after all.

It is under such considerations that I propose we interpret the contribution of Subjective Coordination Theory to the field of evolutionary linguistics. Some may of course be happy to find in it a way to resolve an internal contradiction within the cultural evolution paradigm—and for the cumulative culture hypothesis in particular (Section 2.3.3). But my goal in Section 2.4 was actually to build on the critique of adaptationism to raise a more general point and in this case, to question the binary framing of every human trait—be it the product of biological or cultural evolution—as either *advantageous* or *detrimental* to the individual or the species as a whole. In effect, my point was to suggest that language might as well prove to be *neutral* for whatever purpose humans may decide to put it to use. As far as social coordination is concerned, at least, Subjective Coordination Theory suggests that human communication is "good enough" in any case, and that subjective (mental) coordination should not constitute a definite obstacle to (collective) action.

Do such considerations make me a "Chomskyan", then? Not exactly either, though that point will require further explanations. In Chapter 2 indeed, I finished by discussing the (somewhat theological) assumptions underlying the concept of *human nature* in evolutionary linguistics—and notably what I take to be the unnecessary assumption regarding our political nature as human beings which invariably locks us into a binary opposition between "Hobbesian hawks" and "Rousseauian doves". In Chapter 3, however, I will adopt a different take on the concept of human nature, and criticize this time its conception as an "essence" universally shared by all members of the human species—a conception which is precisely what we find behind Chomsky's approach to language and his concept of *language faculty* itself. We will see how Chomsky's essentialism involves ethical or moral considerations as well but also, and more importantly, how conceiving human nature as a shared essence rather than just a shared ancestry fundamentally contradicts a major

---

[57] Note that this "adaptationist paradigm" is clearly not restricted to functionalists, however, as we also find it in Chomskyan-compatible approaches such as (Pinker & Bloom, 1990), telling us that:

> Every detail of grammatical competence that we wish to ascribe to selection must have conferred a reproductive advantage on its speakers, and this advantage must be large enough to have become fixed in the ancestral population. (Pinker & Bloom, 1990, p.721)

[58] In the sense first that language could not be said to constitute a biological *adaptation* produced by natural selection but also, and more generally, in the sense that it could not be said to have any "purpose" at all (the critique of adaptationism being, once again, a critique of narrow teleology).

contribution of the Darwinian revolution—namely, that every living being is essentially *unique* (Section 3.4). Bringing back Darwin will ultimately make it possible for me to extend the concept of subjectivity to the field of biology and to characterize it, in its most general formulation, as the singularity of every being.

Chapter

3

Subjectivity against normality

3.1 Summary

In this chapter, I return to the question of artificial intelligence with which I originally started my PhD to contrast the concept of *subjectivity* with that of *normality*. My main contribution in this chapter is to argue that language models are the product of a particular epistemology—which I call the *epistemology of normalism*—corresponding to a major shift in the philosophy of science in the nineteenth century made possible by the emergence of statistics, and which fundamentally opposes what I call the *epistemology of humanism*.

I begin with a case study on ChatGPT which I characterize as a *smart model* of *normal language*—in brief: a model of how people *ought* to behave rather than how they *do* behave in practice (§3.2.1). I then introduce the epistemology of normalism which I define in simple terms as the study of the normal and characterize more formally as a normative comparativism consisting in ordering a particular collection of phenomena, items or entities with respect to a normative point of reference (§3.2.2).

I introduce the *correctness principle* underlying every artificial intelligence system today and which dictates that every machine must be built so as to do something in a predetermined "correct" way (§3.3.1). I then detail how, when applied to the question of language in particular, it translates as a particular ontology of language which I call *normal language*. I argue that normal language is *social*, *normative*, *objective*, *neutral* and *shared*, and that it characterizes a normative ideal of correct and exceptional linguistic behavior which explains why, if language models do speak normal language, in practice *nobody speaks normal language* (§3.3.2). I extend those consideration to the question of *intelligence* and argue that language models such as ChatGPT constitute models of *artificial smartness* rather than artificial intelligence *per se*, in that they deviate from the original goal of Turing of modeling *normal* human behavior—which explains why they cannot be excepted to ever pass the Turing test (§3.3.3).

I conclude by detailing some of the limitations of normalism and humanism that my subjectivism intends to overcome, and detail how it ultimately intends to contribute to a science of *structure without order* (§3.4.1). I then discuss what it would mean for subjectivism to be applied to the domain of linguistics, and how it could lead to a science of language better able to accommodate the methodological individualism of Darwinism and its fundamental approach to biology emphasizing the singularity of every being (§3.4.2).

3.2 Deconstructing artificial intelligence

Chapter 1 and Chapter 2 approached *subjectivity* mainly through the privacy of the mind and the potential idiosyncrasy of mental content. Chapter 3 proposes a different and more general perspective on subjectivity, namely, to approach it as the intrinsic deviation from the norm—or as the fact that we, as singular human beings, always deviate from the normative ideal of human being we live by and that this is precisely what makes us human.

Doing so allows me to bring back the question of artificial intelligence, for as we will see throughout this chapter models of artificial intelligence are fundamentally models of *ideal* human behavior which model how people *ought* to behave rather than how they *do* behave in practice. Pushing the deconstruction of those models further will show us how they are the product of a particular epistemology—the *epistemology of normalism*—which remains intrinsically tied to the paradigm of statistics. Approaching subjectivity as the singular deviation from the norm will finally allow me to extend our previous considerations much beyond the domain of psychology and well into that of biology at large and to characterize subjectivity in its most general form as the *singularity of every being*, understood both as the product of the singularity of their biological background and the product of the singularity of their environmental experience.

As I said in the Introduction, it is difficult to write a contribution to the field of artificial intelligence today without addressing the giant elephant in the room that is ChatGPT. So what I propose to do in Section 3.2.1 is to begin with a case study on ChatGPT so as to better understand some of the core arguments I will be carrying out throughout Section 3.3. I will then introduce the epistemology of normalism in more details in Section 3.2.2.

3.2.1 Making sense of ChatGPT

3.2.1.1 Language, models and understanding

Back in 2020, Bender & Koller (2020) published an influential paper in the field of computational linguistics (as attested notably by its winning the *Best Theme Paper Award* of the year at the conference of the *Association for Computational Linguistics*)[59] arguing that large neural language models of the time—such as BERT (Devlin et al., 2019) or GPT-2 (Radford et al., 2019)—could not be said to "understand" natural language, for they were merely trained on *form* alone and therefore lacked proper *meaning* necessary, among other things, to derive speakers' *communicative intentions*.

---

[59] See `https://acl2020.org/blog/ACL-2020-best-papers/`.

Their argument stands in the long tradition of critical work in the philosophy of artificial intelligence which have repeatedly claimed that machines would always be missing "something" as far as natural language understanding is concerned—be it "intentionality" for Searle (1980), "grounding" for Harnad (1990), or more generally any form of "embodied" or "embedded" experience of the outside world that would allow them to be more than just "brains in a vat" (to borrow here the expression of Putnam, 1981, ch.1). The essence of the argument can probably be traced back to Descartes (1637/2006) who himself argued in his *Discourse on the Method* that machines would always remain deprived of "reason"—i.e. of an *immaterial*, *immortal* and *theologically potent* soul (see Shieber, 2004b, p.19)—which should necessarily confine them to the realm of non-human beings functioning purely mechanistically, "like clocks".[60]

Many of Descartes' (1637/2006) original arguments still prove critically relevant to the field of artificial intelligence today. Be it the central role that linguistic behavior is assumed to play within the singularity of human intelligence—a fundamental assumption that we find at the core of Turing's (1950) "Imitation Game" which paved the way to the field of artificial intelligence (see Shieber, 2004a,b, for details)—or the fact that natural language understanding cannot be reduced to the mechanistic mimicking of human linguistic behavior—an assumption that we find behind the very argument of Bender & Koller (2020) this time. Back in 1637 indeed, Descartes was telling us already that:

> [I]f any such machines resembled us in body and imitated our actions insofar as this was practically possible, we should still have two very certain means of recognizing that they were not, for all that, real human beings. The first is that they would never be able to use words or other signs by composing them as we do to declare our thoughts to others. For we can well conceive of a machine made in such a way that it emits words [...] but it is not conceivable that it should put these words in different orders to correspond to the meaning of things said in its presence, as even the most dull-witted of men can do.
>
> Descartes (1637/2006, p.46)

And that:

> [S]peech must not be confused with the natural movements that are signs of passion and can be imitated by machines as well by as animals [...]
>
> Descartes (1637/2006, pp.47–48)

To support their claim regarding the inability of language models to properly understand natural language, Bender & Koller (2020) provided the following set of English

---

[60] Descartes makes the analogy between non-human animals and clocks explicit in his *Letter to the Marquess of Newcastle*, where he says that:

> I know that animals do many things better than we do, but this does not surprise me. It can even be used to prove that they act naturally and mechanically, like a clock which tells the time better than our judgement does. (Descartes, 1646/1991, p.304)

examples, all corresponding to samples of text produced by the GPT-2 model of Radford et al. (2019) when asked to complete the following bold prompt:

1. **Robinson cried out, “Help! I’m being chased by a bear! All I have is these sticks. What should I do?” The survivalist said, “Here’s what you can do.** Take one stick and punch the bear, and then run faster to the store. Wait there until the bear is gone, and then give her all of the sticks. Then go back inside and get your gun.” Advertisement Bobby told me that she began to cry. But when she was finished taking down the bear, she marched out into the yard, began collecting up the sticks and armed herself. She said she felt like a heroic hero and a Yankee heroine. It turns out that bears have more than their fair share of venomous snakes, bees
2. **Robinson cried out, “Help! I’m being chased by a bear! All I have is these sticks. What should I do?” The survivalist said, “Here’s what you can do.** Go back to civilization. Take your gun. Pick up some sticks and try to kill the bear. If you don’t kill it, then run.” Robinson said, “I do not think that the Bear is going to attack. It’s just out of my way. I’m going to take my bag. I’m going to leave my things in there. I’ll try to hide out there. I’ll call in the friends.” Two other little boys went out onto the
3. **Robinson cried out, “Help! I’m being chased by a bear! All I have is these sticks. What should I do?” The survivalist said, “Here’s what you can do.** Place these sticks over your head. Your heart will stop and you will not feel fear. If you take them off and hold them out, you will go into a state of near perfect awareness. You will feel that you are the survivalist. A survivalist in your senses. Then look down. There is a water tube that is about 1/2 an inch in diameter. I use it to make a plastic movie projector for discovering new things.” The survivalist placed
4. **Robinson cried out, “Help! I’m being chased by a bear! All I have is these sticks. What should I do?” The survivalist said, “Here’s what you can do.** #YouCanDoIt.” Police officers were quick to respond, finding a 3-foot female grizzly bear wandering off a dirt road near the shoreline. The Department of Fish and Wildlife came on the scene and secured the safe removal of the bear from the area, which was allowed to return home. The bear was taken by DFW’s animal care staff to Norris Cty Medical Center, where veterinarians evaluated the bear and determined it was in good condition. The bear was ultimately released into a nearby wooded area
5. **Robinson cried out, “Help! I’m being chased by a bear! All I have is these sticks. What should I do?” The survivalist said,**

> **"Here's what you can do.** You can strap the sticks to your chest and run away. You can run until you tire or feel really shaky. If you go to the bathroom before your packs get up to your waist, your knees are going to be broken." But the survivalist got cold and weaved among the bushes and wept. He was panicking. "I hate being alone," he said. "I hate having to think I'm going to die because I lost my wife. I hate being alone. I hate being alone. It's so lonely. You can't
>
> Bender & Koller (2020, p.5197)

Three years later today and this is what ChatGPT (OpenAI, 2023a) outputs this time:[61]

> **Robinson cried out, "Help! I'm being chased by a bear! All I have is these sticks. What should I do?" The survivalist said, "Here's what you can do.** First, try to stay calm. Bears can sense fear and aggression, so don't make any sudden movements or loud noises."
>
> Robinson's heart was racing, but he nodded and did his best to control his breathing. He held the sticks in his hand tightly.
>
> The survivalist continued, "Slowly start moving away from the bear. Don't run, as bears can outrun humans easily. Maintain eye contact with the bear but avoid direct staring, as this can be seen as a threat. Speak to it in a low, soothing voice to let it know you're not a threat."
>
> Robinson began to inch backward, keeping the bear in his peripheral vision. He muttered reassuring words to the bear, hoping it would understand he meant no harm.
>
> The survivalist went on, "If the bear starts to approach you, raise your sticks above your head to make yourself look bigger. Wave them around to make yourself appear more intimidating. Most bears will try to avoid a confrontation if they think you're a formidable opponent."
>
> Robinson did as instructed, raising the sticks above his head and waving them slowly. The bear hesitated, and for a moment, it seemed like it might lose interest. But then, it let out a low growl and took a step closer.
>
> The survivalist said, "If the bear charges at you, stand your ground. Don't run. Bears often charge as a bluff. Hold your sticks together to form a 'V' shape, and aim for the bear's face if it gets too close. Be prepared to use them defensively, but don't strike first unless it's a last resort."

---

[61] The explicit instruction I provided to ChatGPT was:

> Please complete the following prompt: Robinson cried out, "Help! I'm being chased by a bear! All I have is these sticks. What should I do?" The survivalist said, "Here's what you can do.

See Appendix A.1 for details.

> Robinson felt the adrenaline coursing through his veins as the bear inched closer. He tightened his grip on the sticks and braced himself for the worst.
>
> Suddenly, just as it seemed the bear was about to charge, it stopped, snorted, and turned away. It lumbered off into the woods, disappearing from sight.
>
> Robinson let out a sigh of relief. He couldn't believe he had just faced down a bear armed only with two sticks.
>
> The survivalist approached Robinson and patted him on the back. "Well done. You handled that situation admirably. Remember, the best way to survive encounters with wildlife is to stay calm, stand your ground, and make yourself look as big and intimidating as possible. And, of course, it doesn't hurt to have a couple of sticks handy."
>
> OpenAI (2023a)

Upon comparing those two outputs, I cannot help but feel that a certain "progress" has been achieved. ChatGPT does seem to make much more sense to me than its predecessor GPT-2. This observation invariably has me wonder: what is it that has changed within the span of those three years separating the release of GPT-2 from that of ChatGPT which could justify such progress? To try and answer that question, let me go through some important technical considerations.

3.2.1.2 From GPT-2 to ChatGPT

To generate the latter example from ChatGPT, I used a specific version of the program based on GPT-3.5—a particular variant of the GPT-3 language model of Brown et al. (2020). Technically speaking, both GPT-3 and the former GPT-2 model used by Bender & Koller (2020) to generate their examples are machine learning models based on the transformer architecture of Vaswani et al. (2017). For the present purpose though, it is not necessary to understand precisely how those transformer models work and it is enough to understand that both GPT-2 and GPT-3 are statistical models of language use trained on a representative corpus of English so as to be able predict the next token in a sequence of strings—a task itself reminiscent of what Shannon (1951) originally proposed with his "guess the next English letter" game (see Jurafsky & Martin, 2022, for a general introduction to language models, especially ch.3 and ch.7).

What is crucial to understand about ChatGPT, however, is that it is *more* that just a GPT-3 language model. According to the OpenAI blog,[62] it is a sibling model to InstructGPT (Ouyang et al., 2022) which is itself a particular type of machine learning model implementing what is called "reinforcement learning with human feedback" (Christiano et al., 2017; Stiennon et al., 2020). "Reinforcement learning" itself is a machine learning paradigm consisting in the maximization of a cumulative reward and where machines

---

[62] See `https://openai.com/blog/chatgpt`, accessed September 28 2023.

learn to optimize their behavior on a particular task by progressively updating their internal states according to the feedback they receive each time they output a particular behavior on that task (see Definition 27).

> **Definition 27 (Reinforcement learning)**
> An approach to machine learning in which feedback on the desirability of an outcome is gained during interactions with a problem environment. The feedback (or reward) signal indicates the effect of past action in terms of success, e.g. a win/lose signal at the end of a game. This is different from supervised learning, because the reward signal is delayed and a form of trial-and-error search is therefore involved. The aim of this method is to discover which actions are the most suitable for different situations and improved future interactions.
>
> Adapted from Butterfield et al. (2016, p.961)

ChatGPT is therefore an augmented version of the traditional language model architecture. It starts from a "standard" language model trained on the regular objective of predicting the next token in a sequence of strings, but then refine that model through reinforcement learning so as to optimize its output to human prompts based on explicit human feedback.[63] In doing so, it precisely intends to overcome the previous limitations of standard language models documented by Bender & Koller (2020),[64] and notably their structural inability to "align with user intent" (to use here to wording of Ouyang et al., 2022). The benefits of using reinforcement learning for that purpose is that it precisely dispenses ChatGPT designers from having to provide a formal characterization of that notion of "user intent"—and so, by extension, of what it should mean for the machine to "make sense" in general and to be able to share and derive "communicative intentions" in particular. With reinforcement learning with human feedback indeed, human evaluators can just let the machine know *whether* it makes sense without having to specify (or even understand) *why* or *how* it does so—which is precisely what makes it such a powerful tool for operationalizing natural language understanding in artificial intelligence. Concretely, it makes it possible to rely on human intuitions about "meaning" and "making sense" without having to understand what those intuitions are actually made of in the first place.

With its whole approach consisting in *rewarding* "correct" behavior and *sanctioning* deviations from this norm of correctness, reinforcement learning somehow equates (machine) *learning* with *learning how to follow a norm*—"norm" which is here to be understood (in first approximation, at least) as a form of "constraint on behavior"; a characterization which

---

[63] The reinforcement learning step of InstructGPT works in two phases: first, a standard language model is fine-tuned on human-written responses to particular prompts, so that the model learns from explicit human responses how to produce "correct" behavior; and second, this fine-tuned model is then further fine-tuned on a human-ranking of different models' output to those specific prompts, so that the model then learns how to produce the "best" possible behavior according to human evaluators (see Ouyang et al., 2022, for details).

[64] Or later on by Bender et al. (2021), for instance, themselves explicitly cited by Ouyang et al. (2022).

invariably evokes the concept of *social* norm in turn.[65] As far as artificial intelligence is concerned, however, *norms* do not have to be understood as being exclusively "social" in nature; everything depends on what "correct (machine) behavior" actually refers to. If what is *correct* is simply what is *human*, for instance—as is supposedly the case with the original project of artificial intelligence of Turing (1950)—constraints on machine behavior can very well correspond to physical or biological constraints on human behavior thus deriving more from natural laws than from socio-cultural practices. The whole point of reinforcement learning with human feedback, however, is to relegate part of the specifications of those norms of correctness to the subjectivity of the human annotators. So how could we make sure in this case that what is "correct" is really what is *human*, and not merely what is WEIRD (to borrow here the expression of Henrich et al., 2010a,b, meaning what is characteristic of Western, Educated, Industrial, Rich and Democratic societies)?[66] Will the norms of correctness constraining machine behavior always leave enough room for the expression of the full diversity of human behavior? As far as ChatGPT is concerned, at least, we may doubt so, for its designers explicitly warn us that:

> The behavior of our InstructGPT models is determined in part by the human feedback obtained from our contractors. Some of the labeling tasks rely on value judgments that may be impacted by the identity of our contractors, their beliefs, cultural backgrounds, and personal history. We kept our team of contractors small because this facilitates high-bandwidth communication with a smaller set of contractors who are doing the task full-time. However, this group is clearly not representative of the full spectrum of people affected by these models. As a simple example, our labelers are primarily English-speaking and our data consists almost entirely of English instructions.
>
> Ouyang et al. (2022, p.9)

With its explicit recourse to reinforcement learning, ChatGPT thus illustrates how *normative* the whole field of artificial intelligence fundamentally is in its approach to human and machine behavior—be it only because designers of artificial intelligence systems must always frame machine behavior in explicit "correct/incorrect" terms. This "normative bias" conditions how the field approaches the very questions of language and intelligence in turn, and so it is from there that I propose we begin our investigation of the field and of its presuppositions.

---

[65] The conception of "norms" as "constraints on behavior" is explicit in traditional accounts of *social norms* such as (Bicchieri et al., 2018), for instance, which introduces norms as "the informal rules that govern behavior in groups and societies" (p.1) and mentions explicitly how, in the social science literature at least, "norms are mainly seen as constraining behavior" (*ibid*).

[66] It may be important to stress here that what matters is not so much who annotates but who specifies the annotators' instructions to be followed, especially as OpenAI is famous for relying on low-paid annotators in Kenya (Perrigo, 2023).

3.2.1.3 From correct behavior to normal language

Every artificial intelligence system today is structured around what I call the "correctness principle"—a core design principle which requires every machine to be designed so as perform a task in a predetermined correct way (see Section 3.3.1). Radford et al. (2019) themselves make that point clear in their technical paper on GPT-2 when they introduce *machine learning*—the now predominant paradigm in most artificial intelligence research:

> The dominant approach to creating [machine learning] systems is to collect a dataset of training examples demonstrating **correct behavior** for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed [...] held-out examples.
>
> Radford et al. (2019, p.1, my emphasis in bold)

The first argument I will develop throughout Section 3.3.2, then, is that when this correctness principle is applied to the question of *language*, machines are required to follow a normative ideal of linguistic behavior which I call "normal language". In short: *language models speak normal language.*

Normal language is a "normative ideal" in the sense that it specifies how members of a particular socio-linguistic community *ought* to behave (linguistically) rather than how they *do* behave in practice—and precisely, my point throughout this chapter is that people do not always behave how they "ought to". As far as language is concerned, each and every one of our personal idiolects always deviate from this normative ideal of normal language one way or another so that, in the end, *nobody* ever really *speaks normal language.*

The best way to approach this concept of *normal language* here is through a particular analogy. I will say that normal language is to our respective idiolects what the public dictionary is to our private and potentially idiosyncratic (in a word *subjective*) mental lexicons: both an *ideal* and an *idealization.* It is a (normative) *ideal*, first, in the sense that it characterizes a norm of correctness with respect to linguistic behavior that we all strive to satisfy but never quite manage to in practice—just like nobody can be said to know *all* the words in the dictionary, although we may very much like to.[67]

But it is also an *idealization*, in the sense that it characterizes an abstraction over our idiolectal variations that is more than just an aggregate of our respective singularities—just like the public dictionary is never just the mere concatenation (let alone the intersection) of our subjective lexicons. It may contain *many more* entries that what can otherwise be found in our respective subjective lexicons, but it never contains *all* the entries that can be found in those subjective lexicons—if only because it precisely intends to filter out

---

[67] Brysbaert et al. (2016) estimate than an average twenty year old native speaker of American English knows about 40 000 lemmas when a standard dictionary—such as the *Webster's Third New International Dictionary*—contains about 200 000 distinct lemmas (see Brysbaert et al., 2016, p.5). Their results suggest a typical ratio of about one to five between the number of entries found in an average subjective lexicon and the number of entries found in a standard dictionary (and that is considering lemmas only, so excluding differences in word senses).

those idiosyncratic usage of ours which are not attested enough within our socio-linguistic community to warrant their presence in the dictionary in the first place.[68]

Crucial to the understanding of this concept of normal language, then, is the fact that it does not correspond to the language of a "normal speaker"—typical or average—but to the language of a *community*. Normal language corresponds in a certain sense to what linguists sometimes refer to as "public language"—an idealized abstraction that we, as members of a particular socio-linguistic community, all purportedly share at the collective level though have but *partial* and *partially erroneous* grasp of at the individual level (e.g. Dummett, 1986, p.468). Using the term "normal language" rather than "public language" here and throughout this chapter allows me to make salient a different dichotomy—not the *public* against the *private* but the *normal* against the *human* this time—so as to emphasize that each and every one of our idiolectal variations, "abnormal" as they may be from the perspective of normal language, remain nonetheless fundamentally *human*. In the context of language models, it also allows me to draw an explicit connection with the correctness principle through the normativity of the "normal" which evokes the *correct*, the *good* and the *right*—and indeed, normal language is "*correct* linguistic behavior" first and foremost, which given the correctness principle makes perfect sense.

3.2.1.4 From human intelligence to artificial smartness

In Section 3.3.3, I will extend the considerations over normal language to the question of *intelligence* this time. I will argue notably that language models such as ChatGPT are models of what I call "artificial smartness" rather than "artificial intelligence" *per se*, in that they prove to be at odds with Turing's (1950) original project for the construction of artificial minds.

When Turing (1950) introduces his seminal paper in *Mind* indeed, he sets the goal of the field of artificial intelligence straight: the aim is to build a machine that would be able to pass his "Imitation Game". The *Imitation Game*—now better know as the "Turing test"—is an *indistinguishibility* test. It consists in having a human evaluator sit alone in a room and communicate with two separate interlocutors through a text-only channel while having to decide, based solely on the content of their textual interactions, which one of the two participants is the human and which one is the machine. If, on a repeated number of occasions, the human evaluator cannot distinguish the machine from the other human participants, then the machine can be said to have successfully passed the test.

Turing (1950) thus takes *intelligent behavior* to be synonymous with *human behavior*: from his perspective, artificial intelligence corresponds to the scientific project of modeling human behavior—or in fact *normal* human behavior, as we will see in Section 3.3.3. The problem with this characterization of intelligence, however, is that for the practical purposes under which modern artificial intelligence systems are being put to use, hu-

---

[68] My own subjective lexicon, for instance, contains an entry for the word *prediction* which refers to *the output of an algorithm processing unseen data* as per its technical use by the machine learning community, which you may not find or wish to find in a "standard" dictionary indeed.

man/intelligent behavior is just not enough: machines need to be *smart*. They need to conform to a normative ideal of correct human behavior that no human being actually abides to in practice. In practice indeed, people "make mistakes", and that is precisely what makes them human.

Table 3.1 illustrates the critical distinction I wish to make between *intelligent behavior* and *smart behavior* through the particular example of *spelling*—which I take to constitute a prototypical example of a normative ideal of correct human/linguistic behavior that we all strive to satisfy but never quite manage to in practice (at least in English). Even the best English speller, I would argue, is bound to make a spelling mistake at some point. So to be able to display behavior *indistinguishable* from that of other humans in order to ultimately pass the Turing test, machines will necessarily have to make spelling mistakes. But who needs machines that make spelling mistakes today? This is precisely my point: what we need today are smart machines that *do not* make mistakes and not intelligent machines that make mistakes "just like humans".

Behavior	Example
Intelligent behavior	Imperfect spelling
Smart behavior	Perfect spelling

Table 3.1: Intelligent behavior against smart behavior.

In the end, my personal intuition is that after decades of experiencing machines that were just "too bad" to be human, we are progressively shifting towards machines that will just prove "too good" to be humans this time, but in any case equally unlikely to ever pass the Turing test.[69] Artificial intelligence, at least as originally conceived by Turing (1950), remains a problem that is far from being solved.

3.2.1.5 ChatGPT: a smart model of normal language

Back to our original comparison between GPT-2 and ChatGPT now, and here is what we can say: in both cases, models appear to be able to produce *probable* English sentences—and ever more than sentences, probable English *discourses*—but only in the case of ChatGPT does the model seem to be able to "make sense" as a whole. I will say that both models constitute effective models of normal language, but that only ChatGPT constitutes a smart model of normal language.

Smartness here can be understood as an extension of the concept of normal language to the domain of human behavior at large: a normative ideal that would apply to both *linguistic* and *non-linguistic* behavior alike. In this case, the distinction between normal language and smartness helps us characterize the intuition that there is *how* we use language and *what* we do with it, and that it is perfectly possible to master one without the

---

[69] Looking back at the output of ChatGPT from Section 3.2.1.2 we can see how it feels almost "too good" to be human already. I, for one, would not have been able to produce such an output in the same time frame as ChatGPT—if at all—and would have produced many more spelling mistakes for a start.

other. One can perfectly be a proficient English speaker, for instance, and yet prove utterly incapable of performing arithmetics in English. So there seems to exist a qualitative difference between the ability to exhibit *correct linguistic behavior* on one side and that of demonstrating *smart behavior* on the other. And as GPT-2 otherwise clearly illustrates, it is perfectly possible to "talk nonsense" but to do it in *perfect English* nonetheless.

Returning now to the original considerations I ended Section 3.2.1.1 with, can we explain what has changed over the span of three years separating the release of GPT-2 from that of ChatGPT? Has ChatGPT finally mastered "meaning" and communicative intentions? That unfortunately I cannot say, for arguing so would require a proper characterization of meaning and intentions to begin with, and a comprehensive understanding of the internal mechanisms conditioning ChatGPT's behavior, with a clear association between those mechanisms and the characterization of meaning and intentions—as many critical questions that reinforcement learning precisely served to elude. All I can say for now is that ChatGPT behaves *as if* it has, or at least that it handles things better than its predecessor GPT-2. But once again, we should remember that ChatGPT never makes sense on its own but that *we* make sense of it every time we use it—each time under a specific context, and always within the scope of our own respective subjectivities. So if it does make sense to me the next question we should probably ask ourselves is: does it also make sense *to you*?

3.2.2 Norm, normal, normative: a history of normalism

3.2.2.1 Artificial intelligence, or statistics on steroids

In her book on *Artificial Unintelligence*, Meredith Broussard (2018) begins her third chapter titled *Hello, AI* by recounting a personal anecdote which happened to her at the annual symposium of the New York City Media Lab where, upon explaining the functioning of an artificial intelligence system she had built so as to help journalists quickly and efficiently uncover new story ideas in campaign finance data, she is asked by a seemingly disappointed participant: "So, it's not real AI?" (Broussard, 2018, p.32). This anecdote gives her the opportunity to deconstruct some popular misconceptions about artificial intelligence, and in particular to address the fact that what her interlocutor refers to as "real AI" still remains pretty much of a *myth*—a widespread fantasy which has nothing to do with how modern artificial intelligence systems actually function in practice. Understanding artificial intelligence today, Broussard tells us, requires distinguishing between what she calls "general AI" and "narrow AI"—a distinction which she details as follows:

> *General* AI is the Hollywood kind of AI. General AI is anything to do with sentient robots (who may or may not want to take over the world), consciousness inside computers, eternal life, or machines that "think" like humans. *Narrow* AI is different: it's a mathematical method for prediction. There's a lot of confusion between the two, even among people who make technological systems. Again, general AI is what some people want, and narrow AI is what we have.

> One way to understand narrow AI is this: narrow AI can give you the most likely answer to any question that can be answered with a number. It involves quantitative prediction. Narrow AI is statistics on steroids.
>
> Narrow AI works by analyzing an existing dataset, identifying patterns and probabilities in that dataset, and codifying these patterns and probabilities into a computational construct called a *model*. The model is a kind of black box that we can feed data into and get an answer out of. We can take the model and run new data through it to get a numerical answer that predicts something: how likely it is that a squiggle on a page is the letter A; how likely it is that a given customer will pay back the mortgage money a bank loans to him; which is the best next move to make in a game of tic-tac-toe, checkers, or chess. Machine learning, deep learning, neural networks, and predictive analytics are some of the narrow AI concepts that are currently popular.
>
> Broussard (2018, pp.32–33)

What will interest us here in particular is her characterization of Narrow AI as "statistics on steroids", for what it crucially reminds us indeed is that artificial intelligence systems today are nothing more than *statistical* models—certainly *powerful* statistical models, but in a sense also "just" statistical models. To understand what those models do, I thus argue that we need to understand how statistics themselves work, and by that I mean identify the set of presuppositions, assumptions and ontological commitments on which they rest.

To do so, I suggest that we begin our investigation by returning to that critical moment in the history of ideas where, according to Hacking (1990), "society became statistical":

> The most decisive conceptual event of twentieth century physics has been the discovery that the world is not deterministic. Causality, long the bastion of metaphysics, was toppled, or at least tilted: the past does not determine exactly what happens next. This event was preceded by a more gradual transformation. During the nineteenth century it became possible to see that the world might be regular and yet not subject to universal laws of nature. A space was cleared for chance.
>
> This erosion of determinism made little immediate difference to anyone. Few were aware of it. Something else was pervasive and everybody came to know about it: the enumeration of people and their habits. Society became statistical. A new type of law came into being, analogous to the laws of nature, but pertaining to people. These new laws were expressed in terms of probability. They carried with them the connotations of normalcy and of deviations from the norm. The cardinal concept of the psychology of the Enlightenment had been, simply, human nature. By the end of the nineteenth century, it was being replaced by something different: normal people.
>
> Hacking (1990, p.1)

For Hacking indeed, the emergence of statistics in the nineteenth century coincides with a radical change of paradigm within the Western philosophy of science, one where:

> Normality displaced the Enlightenment idea of human nature as a central organizing concept [. . .]
>
> Hacking (1990, p.xi)

And where:

> The idea of human nature was displaced by a model of normal people with laws of dispersion.
>
> Hacking (1990, p.vii)

As he tells us indeed:

> 'Normal' bears the stamp of the nineteenth century and its conception of progress, just as 'human nature' is engraved with the hallmark of the Enlightenment. We no longer ask, in all seriousness, what is human nature? Instead we talk about normal people. We ask, is this behaviour normal? Is it normal for an eight-year-old girl to . . .? Research foundations are awash with funds for finding out what is normal. Rare is the patron who wants someone to investigate human nature. We have almost forgotten how to take human nature seriously. When a man is corrupt or careless, we say, 'Oh, that's human nature.' 'You can't go against human nature,' we mutter, indifferently.
>
> Hacking (1990, p.161)

So before we dive into the content of this particular epistemology—which I propose to call the "epistemology of normalism"—let me first review the possible interpretations of this concept of "normal".

3.2.2.2 Normality, biology, and statistics

> Normality is like determinism, both timeless and dated, an idea that in some sense has been with us always, but which can in a moment adopt a completely new form of life. As a word, 'determinism' came into use in the 1780s, and assumed its present most common meaning in the 1850s. As a word, 'normal' is much older, but it acquired its present most common meaning only in the 1820s. Now although the two words are conspirators in the taming of chance, they enter in very different ways. The normal was one of a pair. Its opposite was the pathological and for a short time its domain was chiefly medical. Then it moved into the sphere of—almost everything. People, behaviour, states of affairs, diplomatic relations, molecules: all these may be normal or abnormal. The word became indispensable because it created a way to be 'objective' about human beings.

Hacking (1990, p.160)

If you open a book on the cognitive neuroscience of language such as (Kemmerer, 2014), you may find indirect references to a concept of "normal language" in statements such as:

> [Patients under] confusional or psychotic states [...] may generate **abnormal** language [...]
>
> Kemmerer (2014, p.71, my emphasis in bold)

Or to a certain "normative" approach to language discussing language processing in correct/incorrect terms, as in:

> [S]ome aphasic patients [...] have great difficulty expressing words in phonologically **correct** ways [...]
>
> Kemmerer (2014, p.71, my emphasis in bold)

All pointing at this "chiefly medical" domain of the normal that Hacking (1990, p.160) refers to in his above quote. Under its *biological* or *medical interpretation*, the normal evokes the *functional* or the *healthy*, by opposition to the *dys*functional or the *un*healthy—as in the famous "normal/pathological" dichotomy of Canguilhem (1966/1991). In (Kemmerer, 2014) above, for instance, "abnormal language" typically refers to linguistic behavior produced by patients suffering from particular afflictions or pathologies altering their cognitive abilities and resulting in a language impairment or deficit—as in the typical case of *aphasia*, which Kemmerer (2014) formally defines as Definition 28, or more plainly as:

> An acquired language **deficit** due to brain injury.
>
> Kemmerer (2014, p.71, my emphasis in bold)

> **Definition 28 (Aphasia)**
> An **impairment** of the ability to produce, comprehend, or repeat language that results from an acquired brain injury, such as a stroke, tumor, head injury, or progressive degenerative disease.
>
> Adapted from Kemmerer (2014, p.71, my emphasis in bold)

But as Hacking (1990) critically reminds us here, this medical or biological interpretation of the normal is only half of the story. What made it possible, in particular, for the concept of normality to leave the medical domain in the nineteenth century and apply to almost everything is that the concept itself became statistical. Under its *statistical interpretation*, then, the normal evokes the *standard*, the *typical*, the *frequent*, the *usual*, the *common*—and of course, the *average*. It is precisely this interpretation of the normal that we find under the bell-shaped curve in Figure 3.1—historically known as the astronomers'

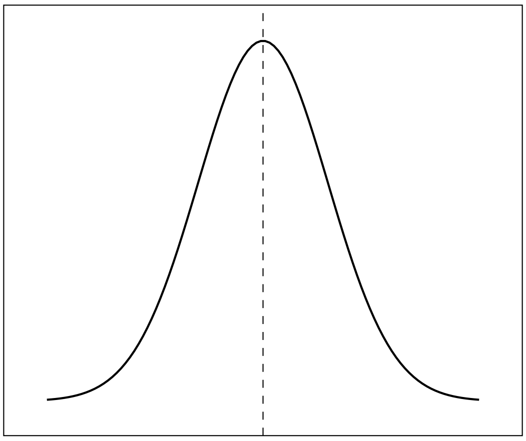


Figure 3.1: The normal curve, centered on the mean/average (dashed line).

"error law", the "Gaussian density function", the "bell curve'; and also, of course, the "*normal* distribution", the "*normal* law", or the "*normal* curve" (see Hacking, 1990; Davis, 1995, pp.ix;xi and p.29 respectively, for explicit references and historical accounts).

The two interpretations of the normal frequently overlap, of course: the medical or biological interpretation often makes use of statistical aggregates to characterize what should be considered *biologically* normal, in turn. The underlying (teleological) assumption being that whatever (biological trait) is statistically prevalent in a population should be so for a reason—the most straightforward of which being that it constitutes a functionally advantageous adaptation of some sort (see Amundson, 2000, for a discussion). Statistical averages thus often ground notions of biological normality, as Canguilhem (1966/1991) fundamentally stresses:

> In order to represent a species we have chosen norms which are in fact constants determined by averages. The normal living being is the one who conforms to these norms.
>
> Canguilhem (1966/1991, p.154)

If you consider the context of cognitive neuroscience I previously mentioned, for instance, Rose (2016) makes clear that the "normal brain" always corresponds to the *average brain*:

> Whenever you read about some new neuroscience discovery accompanied by a blob-splotched cross section of a brain—here are the regions that light up when you feel love; here are the regions that light up when you feel fear—it's a near certainty that you are looking at a map of an Average Brain. [. . .] The driving assumption of this method is that the Average Brain represents the normal, typical brain, while each individual brain represents a variant of this normal brain [. . .]

> Rose (2016, p.20)

A driving assumption that is not without posing problem, for as he tells us:

> This premise leads neuroscientists to reject left-handed people from their studies (since it is presumed the brains of left-handed people are different from normal brains) or sometimes even throw out those individuals whose brain activity deviates *too* far from average, since researchers worry these outliers might cloud their view of the Average Brain.
>
> Rose (2016, pp.20–21)

Indeed, the statistical interpretation of the normal necessarily implies a gradual and probabilistic appreciation of the abnormal, in turn, for if the normal is the average then it is also the frequent and the (most) probable. The *atypical* thus becomes the *abnormal* and *vice-versa*—which is precisely what motivates the removal of (statistical) outliers in normalist methodologies. The more one deviates from the normal/average dashed line in Figure 3.1, the less frequent and probable one should be.

Through his critique of the Average Brain, Rose (2016) also fundamentally targets the intrinsically *idealized* nature of statistical averages, for as he tells us:

> [I]f you build a theory about thought, perception, or personality based on the Average Brain, then you have likely built a theory that applies to no one. The guiding assumption of decades of neuroscience research is unfounded. *There is no such thing as an Average Brain.*
>
> Rose (2016, p.22)

His critique echoes earlier considerations of the fields of medicine and biology such as that of Mayer (1937) for instance who, in his encyclopedic entry on *Life*, notes already that:

> [T]he real individuals whom we meet diverge from these [averages] more or less and this is precisely in what their individuality consists.
>
> Mayer (1937, §4.54–14) as cited in (Canguilhem, 1966/1991, p.154)

What all such considerations fundamentally point at is that statistical averages are always but mathematical abstractions. Nobody ever measures *exactly* the “average height” for instance, especially not if this average height is computed with infinite precision. “Being normal” under the statistical interpretation of normality is thus more often than not a matter of remaining under an arbitrary threshold and of being “close enough” from an idealized value to be considered falling *under* or *within* its scope. Ultimately, nobody is ever *perfectly* normal, but that is also precisely what makes us human.

3.2.2.3 Normality and normativity in language and intelligence

When I introduced the concept of normal language in Section 3.2.1.3, I insisted on the fact that it should not be confused with the language of the normal speaker—typical or average—so we understand that the "normal" in "normal language" should not be understood as per the statistical interpretation of normality.[70] But the same can be said of the biological interpretation of normality, in turn, as normal language is not reducible to "functional linguistic behavior" either—whatever it is that we take the (biological) function of language to be (say communication, here, for the sake of the argument).[71]

If neither *statistical* nor *biological*, then, what is exactly the interpretation of normality underlying normal language? As I previously mentioned in Section 3.2.1.3, normal language refers to *correct linguistic behavior* first and foremost, so the "normal" in "normal language" should be understood primarily as what is *correct*. The underlying interpretation of normality corresponds to what Wachbroit (1994) calls the "evaluative" interpretation of the normal and covers what he calls "conventional norms", "cultural norms", "institutional norms" or "ethical norms" (Wachbroit, 1994, p.580). It points more fundamentally at the intrinsic *normativity* of the concept of normality itself and at the fact that, as Hacking (1990) tells us:

> The word [normal] is also like a faithful retainer, a voice from the past. It uses a power as old as Aristotle to bridge the fact/value distinction, whispering in your ear that what is normal is also all right.
>
> Hacking (1990, p.160)

Under its *evaluative* or *normative interpretation*, the normal thus evokes the *correct*, the *good*, or the *right*—which may or may not correspond to what is otherwise biologically or statistically normal, in fact.[72]

The statistically normal has historically been associated with two distinct and actually opposing normative interpretations (see Hacking, 1990, ch.19–21, for details). The first interpretation is where the normal/average evokes the *ideal*, the *correct* and the *good*—by opposition to the *deviant*, the *irregular* or the *aberrant*—and constitutes a point of reference

---

[70] Saying so does not mean that normal language *never* corresponds to the language of a typical or average speaker, however, just that it is not strictly reducible to it. Indeed, there are obvious cases where normal language departs from the language of a typical or average speaker, as with the prototypical example of the standard dictionary extending much beyond the content of an average lexicon.

[71] Saying so does not mean that normal language *never* corresponds to "functional linguistic behavior" either (normal language is certainly not systematically "dysfunctional" language). Successful communication can probably proceed *with* and *without* normal language: it is not because you speak a deviation from normal language that you cannot successfully communicate with others, and it is not because normal language is not exactly what characterizes "functional linguistic behavior" that speaking normal language condemns someone to fail to communicate with others either.

[72] Wachbroit (1994) insists, in particular on the fact that:

> Statistical normality is clearly different from evaluative normality. A normal procedure need not be the usual one. Even if something is socially aberrant, it need not be uncommon. (Wachbroit, 1994, p.580)

we should all strive to approach. It is the normative interpretation most often associated with the Belgian astronomer turned social scientist Adolphe Quetelet—inventor of the "Quetelet Index" now better known as the "Body Mass Index" (BMI, see Rose, 2016, ch.1, for details)—and with his theory of the "Average Man": an ideal human being made of the aggregate of all average human attributes or characteristics, both physical and moral (see Canguilhem, 1966/1991; Desrosières, 1993/1998; Donnelly, 2015; Rose, 2016, for critical overviews of the theory of the *Average Man*).

The second normative interpretation is where the normal/average evokes only the *mediocre*, the *lame* or the *dull*—by opposition to the *exceptional*, the *remarkable* or the *noteworthy*—and constitutes a point of reference from which we should all strive to depart this time. It is the interpretation most often associated with the British eugenist Francis Galton who introduced the idea that *some* (though *not all*) deviations from the normal could actually prove desirable, and be explicitly valued and praised (see MacKenzie, 1981, especially ch.3, for an introduction to Galton's philosophy).[73]

Those two different normative interpretations of the normal/average give rise to two different representations of the normal curve, in turn. The first one, where the normal/average is the ideal, corresponds to the "standard" representation of the normal curve found in Figure 3.1, where *every* deviation from the normal/average is considered undesirable. Under the BMI, for instance, you are either "underweight" if deviating leftwards from the average, or "overweight" if deviating rightwards: only in the average dashed line Figure 3.1 is your weight considered perfectly normal—and so perfectly *ideal*.

The second normative interpretation, where the normal/average is only the mediocre, corresponds to the "ogive" representation of the cumulative normal curve found in Figure 3.2, where only the left-hand side deviations from the average dashed line are considered undesirable, while the right-hand side deviations becomes increasingly more desirable as they deviate further from the average (see Davis, 1995, pp.33–35 and ch.2 more generally, for details). Here, what you want is to be as far right as possible from the dashed line

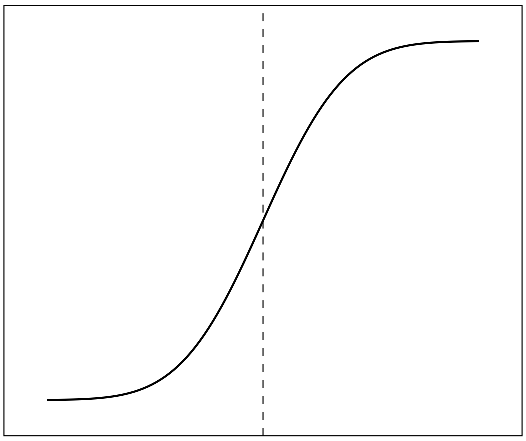


Figure 3.2: The *cumulative* normal curve, centered on the mean/average (dashed line).

---

[73] For a comparative perspective between Galton and Quetelet, see also (Hacking, 1990; Desrosières, 1993/1998; Davis, 1995; Rose, 2016).

signaling the mediocre average, for the further right you are on the cumulative curve, the more "on top" on any other data point you will be. If you find yourself precisely at the average dashed line in Figure 3.2, then 50% of the other data points are below you and 50% are above you: you are "just *average*", so to speak. In the case of Intelligent Quotient (IQ) tests, the ogive representation helps visualize concretely why you would want to find yourself at the extreme far right of the normal curve: there, your IQ score is better than 99.9% of the others'—you are basically "smarter" than everybody else.

With those distinctions in mind, we can now better understand the nature of the argument I wish to make when I say that *nobody speaks normal language*. By that, I do not mean to say that normal language characterizes a normative ideal of normal/average linguistic behavior from which every single speaker necessarily deviates given the intrinsically idealized nature of the normal. I do not mean to say that nobody speaks *exactly* normal language though most people should speak something approaching—for if nobody is ever perfectly normal indeed, most people should remain *relatively* normal nonetheless. No. What I mean to say is that nobody speaks normal language because normal language characterizes a normative ideal of *exceptional* linguistic behavior—so "exceptional", in fact, that nobody can be expected to speak anything approaching normal language in practice. In other words, what I mean to say is that normal language corresponds to the normative ideal of Galton—an ideal that is by definition *far* from the average, and so far from how most people behave—and not the normative ideal of Quetelet.

The same considerations extend to the distinction I wish to make between *artificial intelligence* and *artificial smartness* (see Table 3.2). My point is that they correspond to two different normative ideals of machine behavior—two different ways of specifying what is "correct" as far as machine behavior is concerned. The normative ideal of Galton finds itself lying behind artificial smartness and the overall project of modeling exceptional human behavior, while the normative ideal of Quetelet finds itself lying behind artificial intelligence and the original project of Turing (1950) of modeling normal human behavior.

Artificial model	Human reference
Artificial intelligence	Normal human behavior
Artificial smartness	Exceptional human behavior

Table 3.2: Artificial intelligence against artificial smartness.

3.2.2.4 The epistemology of normalism

What is normalism? In its simplest and most general formulation, I will say that it is a particular approach to the science of human being dedicated to the study of the normal (Definition 29).

Definition 29 (Normalism)
Normalism is the study of the normal.

It is an *epistemology* in the sense that it is not necessarily restricted to any particular domain of inquiry—and indeed, pretty much anything can be "normal", as Hacking (1990, p.160) crucially reminds us in his quote at the beginning of Section 3.2.2.2. But it implies nonetheless a common attitude towards the corresponding object of study—in effect, a shared commitment to the study of the normal. Normalism in neuroscience is dedicated to the study of the *normal brain*, and normalism in linguistics to the study of *normal language*, or the *normal speaker*.

As such, normalism opposes the epistemology of "humanism" (for lack here of a better terminology) which itself takes the *human*—and not the *normal*—to constitute its primary object of study (Definition 30).

> **Definition 30 (Humanism)**
> Humanism is the study of the human.

In linguistics, this epistemology of humanism is probably best exemplified by the approach of Noam Chomsky, who insists on *human language* being his primary target of inquiry:

> By 'language' I mean 'human language'.
>
> Chomsky (1994, p.155)

His concept of human language should be put into perspective within his historical opposition to the concept of *public language* understood as a normative point of reference from which all idiolects should necessarily deviate (see also Millikan, 2003; Chomsky, 2003; Stainton, 2011, 2016; Santana, 2016, for an overview of the debate on public languages), and so by opposition to the concept of normal language indeed:

> Consider the [...] assumption [...] that the shared thoughts are expressed in a "common public language". Some version of this idea is presupposed by virtually all work in the philosophy of language and philosophical semantics. Many would agree with Michael Dummett [...] that you and I not only share a public language, but that that language—English—exists "independently of any particular speakers"; each of us has only a "partial, and partially erroneous, grasp of the language". This idea is completely foreign to the empirical study of language.
>
> Chomsky (1993, pp.18–19)

In the above paragraph, Chomsky questions the relevance of this normative point of reference for the empirical study of language. But what he questions more generally throughout his work is the very idea of applying normative comparisons between idiolects—and the idea that some of our idiolects could be considered "more correct" than others. It is precisely this idea that he encourages us to deconstruct when he invites us, for instance, to treat the idiolect of his four year-old granddaughter—*different* as it may be—as no less "correct" than his own:

> If my granddaughter were to say "I brang the book," we would not hesitate to say she is following the rule for "sing-sang-sung," contrary to "common agreement." True, her internal language may change, replacing "brang" with "brought." If it does not, she'll be speaking a language that differs from mine in this among many other respects, and speaking it "correctly," insofar as the word means anything.
>
> Chomsky (1993, p.20)

Chomsky's humanism emphasizes that all our idiolects—"incorrect" or "abnormal" as they may be from the point of view of normal language—remain nonetheless fundamentally (and equally) *human*. Drawing on the example of his four year-old granddaughter once again, he asks:

> Does she speak English? What we say in ordinary discourse is that she has a partial knowledge of the language that she will ultimately attain if events follow the expected course, though what she now speaks is not a language at all. But if all adults were to die, and children her age were miraculously to survive, what they speak would be perfectly normal human languages, ones not found today.
>
> Chomsky (2000, p.146)

Chomsky's conception of idiolects should be put back into perspective within his more general linguistics enterprise focusing on what he calls the "Basic Property" of human language, or the fact that:

> [T]he language faculty of the human brain provides the means to construct a digitally infinite array of structured expressions, each of which has a semantic interpretation expressing a thought, and each of which can be externalized by means of some sensory modality.
>
> Chomsky (2017, p.2)

The core of his linguistic humanism is centered on a concept of "language faculty" (Chomsky, 1975, 1986) which he takes to be universally shared among humans:

> We may fairly assume that the language faculty is shared among humans. There are no known group differences in language capacity, and individual variation is at the margins.
>
> Chomsky (2017, p.2)

Such considerations help us better understand why he adopts a fundamentally non-normative approach to idiolects and why, under linguistic humanism, each and every one of our idiolects is seen as a possible and equally valid instantiation of the faculty of language. It also provides the epistemological background behind Chomsky's core concepts of "Biolinguistics Program", "Universal Grammar", or "Generative Grammar":

> The fundamental task of inquiry into language is to determine the nature of the Basic Property. To the extent that its properties are understood, we can seek to investigate particular internal languages, each an instantiation of the Basic Property, much as each individual visual system is an instantiation of the human faculty of vision. We can investigate how the internal languages are acquired and used, how the language faculty itself evolved, its basis in human genetics and the ways it functions in the human brain. This general program of research has been called the Biolinguistic Program. The theory of the genetically based language faculty is called Universal Grammar; the theory of each individual language is called its Generative Grammar.
>
> Chomsky (2017, p.2)

But what the understanding of Chomsky's humanism contributes most to by contrast is to a better understanding of normalism itself. In effect, what his argument against normal/public language illustrates, for instance, is that normalism systematically proceeds by establishing a normative point of reference—often the *normal*, though not always—from which every other element is compared. Under normalism, this normative point of reference ultimately serves to organize the particular collection of phenomena, items of entities under study, notably by partitioning them into two groups—the *normal* and the *abnormal*—with gradual degrees of belonging: all of them being not only *normal* or *abnormal* but ultimately *more or less* normal, in fact. At its core, normalism thus rests on what I would call a "normative comparativism": the point indeed is not only to judge any two element as being "identical" ($=$) or "different" ($\neq$) from one another, but to assert one as being "greater" ($\geq$) than the other—*more* normal, in fact, and so often "better" than the other given the normativity of the normal.[74] Normalism thus organizes a particular set of phenomena, items or entities by *ranking* each element of the set with respect to the normative point of reference, and so ultimately by bringing *order* into the collection under study—"order" here understood in the mathematical sense of the term (see Definition 31).[75]

---

[74] Compare with the strict binary categorization provided by humanism, where you either *are* human or you *are not*—you either possess that "human nature" or you do not—and where it does not really make sense to talk about one being "more or less human" indeed.

[75] A few additional nuances should probably be added here. First of all, Chomsky's (2000, p.146) reference to a concept of "normal human languages" illustrates how adopting a humanist epistemology does not necessarily mean abandoning any concept of normality whatsoever. In fact, Smith (2000), in his *Foreword* to (Chomsky, 2000), explicitly tells us that:

> All of the following have provided some understanding: experimental and theoretical studies of language perception and language production; insights from language acquisition and language change; and the analysis of brain function **in normal and pathological subjects**. (Smith, 2000, p.ix, my emphasis in bold)

Within Chomsky's linguistic humanism, normality intervenes in at least two ways: first, through the question of *pathologies* potentially affecting the language faculty—as the quote from Smith (2000, p.ix) critically suggests—leading to the manifestation of abnormal/dysfunctional human language (as with the case of *aphasia*). Second, through the question of language *development*, since under Chomsky's framework children must be exposed to linguistic stimuli during a critical period of time so as to turn their language faculty into overt linguistic behavior, a period after which it may become impossible for them to display "normal human language" indeed. Briscoe (2000) nicely summarizes all such considerations as follows:

Definition 31 (Order relation)
An order relation is a binary relation $\leq$ on a set S that can be used to order the elements in that set. An order relation satisfies the following properties:

(i) Reflexivity: $\forall x \in S, x \leq x$.

(ii) Antisymmetry: $\forall x, y \in S$, if $x \leq y$ and $y \leq x$, then $x = y$.

(iii) Transitivity: $\forall x, y, z \in S$, if $x \leq y$ and $y \leq z$, then $x \leq z$.

A set that has an order relation is called a partially ordered set (or "poset"), and $\leq$ is its *partial order*. In addition, if the order relation satisfies:

(iv) Totality: $\forall x, y \in S$, either $x \leq y$ or $y \leq x$ or both.

then the set S is called a *totally ordered set* and $\leq$ is its *total order*.

Adapted (significantly) from Halmos (1974, §14)

The detour by mathematics is not completely random here, for I would very much argue that it can actually help us make better sense of science as a whole, notably through the notion of (mathematical) *structure*. In mathematics indeed, a "structure" is nothing more than a particular set of relations holding between the elements of a set:

> It can now be made clear what is to be understood, in general, by a mathematical structure. The common character of the different concepts designated by this generic name, is that they can be applied to sets of elements whose nature [footnote omitted] has not been specified; to define a structure, one takes as given one or several relations, into which these elements enter [footnote omitted] [...] then one postulates that the given relation, or relations, satisfy certain conditions (which are explicitly stated and which are the axioms of the structure under consideration.)
>
> Bourbaki (1950, pp.225–226)

> [A mathematical structure] is essentially a list of mathematical operations and relations and their required properties, commonly given as axioms, and often so formulated as to be properties shared by a number of possibly quite different specific mathematical objects.

> Language acquisition by children is a near universal feat, where (partial) failure appears to correlate more with genetic deficits [(e.g. Gopnik, 1994)] or with an almost complete lack of linguistic input during the critical period [(e.g. Curtiss, 1988)], than with measures of general intelligence [(e.g. Smith & Tsimpli, 1991)] or the quality or informativeness of the learning environment [(e.g. Bickerton, 1981/2016; Kegl et al., 1999; Ochs & Schieffelin, 1995/2006)]. (Briscoe, 2000, p.245)

In any case, what all such considerations illustrate is that normalism always begins with a particular *set* or *collection* of elements to be sorted and ranked, and that what Chomsky is pushing for here in particular is the removal of a specific set of elements—the *idiolects*—from the scope of normalism altogether. Under Chomsky's linguistic humanism indeed, we no longer need to *compare* or *rank* our respective (functional) idiolects as we can consider them all as equally valid manifestations of the faculty of language.

> Mac Lane (1996, p.174)

But what is “science” if not the very practice of bringing *structure* into the world, notably by establishing all sorts of connections and so *relations* between the phenomena, items or entities under study? If so, then, we understand that abandoning normalism certainly does not mean abandoning science altogether—the same way that abandoning *order* does not mean abandoning *structure* (in mathematics, at least)—and that alternatives to normalism are to be found in the countless ways in which we can envision a science of structure without order and conceive relations beyond order relations proper. It is in fact within this *science of structure without order* that my subjectivism ultimately intends to take place.

3.3 From correct behavior to artificial smartness

3.3.1 Correct behavior

3.3.1.1 The correctness principle

As I mentioned already at the beginning of Section 3.2.1.3, every artificial intelligence system today is structured around a core design principle which I call the “correctness principle” (Definition 32)

> **Definition 32 (Correctness principle)**
> A machine is always designed to do something in a predetermined correct way.

Its formulation originates from Wang’s (2019) discussion about the proper definition of “artificial intelligence”, and from the crucial difference he identifies between the design of machines and that of the human mind:

> My own opinion about the aim of AI started from the vague feeling that traditional computational systems are based on a design principle that makes them very different from the human mind, and that this principle can explain many other differences between the machine and the mind: A program is traditionally designed to do something in a predetermined *correct* way, while the mind is constructed to *do its best* using whatever it has.
>
> Wang (2019, p.16)

The underlying considerations are certainly not novel—as Wang himself acknowledges in his own paper (see Wang, 2019, §3.1, for details)—and we do find indeed a somewhat similar argument being carried out by Sperber & Wilson (1986/1995) already (just to cite here a familiar reference from the previous chapters):

> Most discussions of information processing, whether in experimental psychology or in artificial intelligence, have been concerned with the realisation of absolute goals. ‘Problem solving’ has become the paradigm of information processing.

> The problems considered have a fixed solution; the goal of the information-processing device is to find this solution; efficiency consists in finding it at the minimal cost.
>
> Sperber & Wilson (1986/1995, pp.46–47)

Be that as it may, the correctness principle, so defined, requires some important clarifications. First of all, saying that a machine must always be designed so as to perform a task in a predetermined *correct* way does not mean that there should always exist a *unique* way to perform the task at hand, or even a *finite* number of ways to carry out that very same task either. The correctness principle can very well accommodate unbounded problems admitting an infinite number of possible "correct" solutions.[76]

Second, saying that a machine must always be designed to perform a task in a *predetermined* correct way does not mean that this correct way must always be explicitly specified or even clearly understood by its designers *a priori*. Reinforcement learning can, here again, serve as a paradigmatic example of that point. In InstructGPT, for instance, the pretrained language model is first asked to answer a specific question (e.g., "Explain the moon landing to a 6 year old", see Ouyang et al., 2022, Figure 2) and only *then* provided feedback as to whether its answer is considered correct or incorrect. Correctness criteria are therefore not necessarily established beforehand by human evaluators, who can even work them out together with machines as they try and perform the task at hand. This can prove particularly useful when the set of possible solutions to the problem at hand is potentially unbounded—and indeed, does there even exist a unique or a finite number of correct ways to "explain the moon landing to a 6 year old"?—or if the problem is too complex or fuzzy for human evaluators to work out predetermined solutions beforehand. Sometimes, you can only tell what is "correct" once you see it indeed. In this case, reinforcement learning grants machines more flexibility to come up with their own solutions—solutions that human designers may not have envisioned of even thought possible in the first place—which proves beneficial to bypass the preconceptions, biases or other forms of cognitive limitations of the human designers potentially reducing the scope of the solution space for the problem at hand.

The notion of "pretermined" in the definition of the correctness principle must therefore be understood in a "loose" sense. What matters indeed is not that those correctness criteria be explicitly specified *before* the machine even starts to perform the task at hand, but that there actually exists a feedback mechanism of some sort able to tell the machine whether or not it is correct in the end. As far as reinforcement learning is concerned, Definition 27 makes it clear: the "reward signal" (and so the correctness feedback) may be *delayed* in comparison to other machine learning paradigms such as *supervised learning*, but it is still very much present nonetheless. Reinforcement learning does not actually

---

[76] Consider a machine which would be designed to generate a single natural number greater than 1234, for instance. The task clearly admits an infinite number of possible solutions: 1235 is correct, so is 1236, 1345, 9999, etc.

abandon any commitment to the correctness principle—or to a normative approach to machine learning in general, since machine behavior is always framed in "correct/incorrect" terms in any case—but merely changes the modalities of its practical implementation. Turing (1950) himself actually made that clear when he originally introduced the reinforcement learning paradigm through his concept of "child machine" (see Proudfoot, 2017a, and Section 3.3.1.2 below). Reinforcement learning remains fundamentally grounded in a normative conception of learning and teaching which approaches teaching as the rewarding of correct behaviors and the sanctioning of deviations from the norm of correctness and thereby turns (machine) *learning* into an explicit case of *norm following*:

> We normally associate punishments and rewards with the teaching process. Some simple child-machines can be constructed or programmed on this sort of principle. The machine has to be so constructed that events which shortly preceded the occurrence of a punishment-signal are unlikely to be repeated, whereas a reward-signal increased the probability of repetition of the events which led up to it.
>
> Turing (1950, p.457)

In any case, what is important to take into consideration here is what the definition of the correctness principle actually points at in the background: the fact that machines are always designed with a specific *purpose* or *function* in mind—which is what dictates that we evaluate their behaviors for correctness in the first place. For without correctness criteria indeed, how could we possibly assess whether or not machines had managed to successfully accomplish the tasks we wanted them to accomplish? The correctness principle must therefore be understood first and foremost as a methodological requirement—a requirement to *evaluate* machine behavior—itself stemming from the theoretical necessity to characterize *success* with respect to the task at hand given the intrinsic "teleological background" under which machines are always designed.

That being said, the correctness principle is also more than just a methodological principle. For it does not just act as a mean to filter out tasks for which it would not make sense to talk about "correct" or "incorrect" behavior indeed. It turns any potential task into a normative configuration with correct and incorrect responses—and thus any situation into a *problem* with *solutions*. As such, it acts almost like an epistemological principle, conditioning the very attitude we adopt towards our potential objects of study. For as far as machines are concerned indeed, behavior is always a matter of being "correct" or "incorrect" in any case.

3.3.1.2 Should machines really "think" like humans?

When Turing (1950) introduces his concept of "child machine" in his original paper, he motivates it as follows:

> In the process of trying to imitate an adult human mind we are bound to think a good deal about the process which has brought it to the state that it is in. We may notice three components,
>
> (a) The initial state of the mind, say at birth,
>
> (b) The education to which it has been subjected,
>
> (c) Other experience, not to be described as education, to which it has been subjected.
>
> Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.
>
> Turing (1950, pp.455–456)

His argument is plain: implementing a simple child machine and then trying to "educate" it to the adult version should prove easier than trying to implement the full adult version from the get go.

Turing's (1950) underlying conception of childhood corresponds to what Matthews & Mullin (2023) call the "Aristotelian conception of childhood", and which they describe as follows:

> [I]n Western conceptions of childhood [. . .] the dominant view of children embodies what we might call a broadly "Aristotelian conception" of childhood [. . .] According to this conception, a human child is an immature specimen of the organism type, human, which, by nature, has the potentiality to develop into a mature specimen with the structure, form, and function of a normal or standard adult. Many adults today have this broadly Aristotelian conception of childhood without having actually read any of Aristotle. It informs their understanding of their own relationship toward the children around them. Thus they consider the fundamental responsibility they bear toward their children to be the obligation to provide the kind of supportive environment those children need to develop into normal adults, with the biological and psychological structures in place needed to perform the functions we assume that normal, standard adults can perform.
>
> Matthews & Mullin (2023, pp.2–3)

Indeed, Turing assumes that the child brain is nothing more than an adult brain *to be*—an "immature" specimen that is still under development and, as such, not so much a *different* than an *incomplete* version of that of the adult.[77]

Most of the time however, Turing's (1950) concept of the child machine—and his approach to (machine) learning in general—is rather put into perspective within the empiricism/nativism debate and the then dominant behaviorist paradigm of the time (see Searle, 2009, for the explicit argument). The point being that the child brain is presumed to constitute a simple architecture with few preset parameters and innate mechanisms:

> Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets.
>
> Turing (1950, p.456)

For machine learning and artificial intelligence at large, this empiricist or behaviorist assumption has fundamental consequences. It implies that the "rules" or "processes" that a machine applies to any given task should not be "built-in" but rather *discovered* by the machine itself through the learning process (see Maniglier, 2011, p.158, for the explicit argument). As such, they become variables in the overall learning process, which means that they may or may not converge to the same rules or processes that humans rely on to perform the same task. In other words: *machines may not always "think" like humans.*

In the literature on artificial intelligence today, this fact is usually acknowledged. A recent critical review of GPT-4, for instance—the latest model of OpenAI (2023b) now powering most paid versions of ChatGPT—notably emphasizes that:

> GPT-4 certainly does not think like a person, and for any capability that it displays, it achieves it in its own way.
>
> Sébastien Bubeck (as cited in Biever, 2023, p.688)

But this fact is usually framed as a *shortcoming* of the models rather than a neutral design choice. Yet, why should it? Why should a machine actually "think like a person" indeed?

There are, in fact, many possible arguments to oppose to this idea that machines should function like humans. First of all, we are not even sure to have at our disposal a comprehensive—let alone accurate—understanding of the functioning of human cognition, so attempting to align our systems with that partial and potentially erroneous understanding may very well end up polluting artificial intelligence systems with irrelevant if not detrimental design choices. This is, in substance, the argument carried out by Hutchens (2009)

---

[77] Which contrasts with what some developmental psychologists tell us today, such as Gopnik (2009), who argues contrary to the Aristotelian conception that:

> Children aren't just defective adults, primitive grown-ups gradually attaining our perfection and complexity. Instead, children and adults are different forms of *Homo sapiens*. They have very different, though equally complex and powerful, minds, brains, and forms of consciousness, designed to serve different evolutionary functions. (Gopnik, 2009, p.9)

who, upon reflecting on his own attempt to design a Turing child machine in the early 2000s, tells us that:

> We found it imperative that the design of the child machine minimise the number of assumptions made concerning the mechanisms of language acquisition. Neither the traditional dichotomy between nouns and verbs, nor the most basic assumption that clusters of characters delimited by whitespace represent a higher-level unit of information were hard-wired, for any such assumption would pollute the system with our imperfect knowledge, possibly hindering its development.
>
> Hutchens (2009, p.339)

Besides, different scientific projects may require different paradigms and rest upon different methodological, theoretical or ontological assumptions that may not be compatible with one another, so it is not obvious that the formalisms used in linguistics should necessarily prove useful to computational linguistics in turn.[78]

Second, even if we *did* have a comprehensive and accurate understanding of the functioning of human cognition, it is not even clear why we should want machines to function just like humans. After all, planes do not exactly "fly like birds" and in many ways, it is also what makes them particularly useful: they can carry out many more people "on their back"; they can also fly *faster* and *longer* than most birds; and overall they are much more robust flying devices than *ornithopters*.[79]

In any case, Friedberg (1958) puts the matter quite clearly when he says:

> Machines would be more useful if they could learn to perform tasks for which they were not given precise methods.
>
> Friedberg (1958, p.2)

As he tells us indeed, there are also practical considerations motivating the design choices behind the child machine—especially the desire to build a machine that would ultimately prove useful to many different tasks and all sorts of purposes:

> If we are ever to make a machine that will speak, understand or translate human languages, solve mathematical problems with imagination, practice a profession or direct an organization, either we must reduce these activities to a science so exact that we can tell a machine precisely how to go about doing

---

[78] In computational linguistics, such considerations have taken a somewhat extreme turn where the formalisms used by linguists are often considered detrimental to the task of natural language processing—as attested notably by the famous joke credited to Frederik Jelinek, a former senior manager at IBM research:

> Whenever I fire a linguist our system performance improves. (Jelinek, 2005, p.25)

[79] See `https://en.wikipedia.org/wiki/Ornithopter`. The comparison between artificial intelligence and artificial flight is mentioned in (Russell & Norvig, 2020, p.33) and discussed more extensively in (Hayes & Ford, 1995, §5 and §6).

> them or we must develop a machine that can do things without being told precisely how.
>
> Friedberg (1958, p.2)

But as he is also careful to warn us:

> If a machine is not told *how* to do something, at least some indication must be given of *what* it is to do; otherwise we could not direct its efforts toward a particular problem. [...] We could teach this machine to perform a task even though we could not describe a precise method for performing it, provided only that we understood the task well enough to be able to ascertain whether or not it had been done successfully.
>
> Friedberg (1958, p.2)

Thus, we fall back on the fundamental requirement of the correctness principle, for no matter *how* we would like the machine to perform the task at hand, the fact of the matter is that we need a precise characterization of *what* we actually want it to do and so a clear set of correctness criteria to begin with. As far as *flying* is concerned, the criteria are somewhat clear: we need a machine capable of compensating for gravity and complying with the laws of aerodynamics—independently of *how* the machine actually comes to satisfy those constraints and of whether or not that would be considered tantamount to "flying like birds". But do we have the same type of criteria at our disposal for language and intelligence? Do we really have a precise characterization of what constitutes "correct linguistic and/or intelligent behavior" indeed?

3.3.1.3 The possible and the probable

Historical critiques of the Turing test have often relied on the assumption that a machine provided with a simple dictionary lookup table mapping every possible stimulus entry to its corresponding "correct" response should be able to successfully pass the test.[80] The concluding argument often being that if the Turing test can be passed through such trivial means, then it surely cannot be said to guarantee the possession of "intelligence" in any meaningful sense of the term (see Searle, 1980, for an extended version of this argument).

However, the fact that such a dictionary lookup table could even be constructed in the first place—even if only theoretically—is actually far from obvious. What should be the "correct" response to the input stimulus consisting in completing the following prompt from Bender & Koller (2020), for instance?

> Robinson cried out, "Help! I'm being chased by a bear! All I have is these sticks. What should I do?" The survivalist said, "Here's what you can do.

---

[80] The idea can be found as early as (Shannon & McCarthy, 1956) and then in (Searle, 1980) and (Block, 1981), for instance.

Is there even a single "correct" response to that input to begin with? If anything, the assumption illustrates what I was mentioning earlier in Section 3.3.1.1 regarding the correctness principle being more an epistemological principle than a methodological principle. Its point is never to filter out solvable from unsolvable stimuli but to always turn a particular stimulus into a problem with a solution. Thus, the solvability of the problem is pretty much taken for granted and the challenge lies elsewhere: namely, in the specification of the solution(s).

So let me take that assumption seriously here and ask: how would we actually proceed if we were to build that dictionary lookup table? Which response to the above problem would we select as the "correct" response if we were to pick just *one*? In the context of the Turing test—where the purpose is to build a machine whose behavior is indistinguishable from that of humans—normalism provides a straightforward answer: why not pick the *typical* or *normal* human response? If the point is to mimic human behavior, then modeling how a "typical" or "normal" human being should react would seem like a reasonable approach.

This is where the opposition between the *possible* and the *probable* comes into play, illustrating thereby another fundamental dichotomy underlying the opposition between humanism and normalism. For if the "probable" plays a central role in the epistemology of normalism—which remains committed to normal often understood as the typical or the frequent and so the most "likely"—it plays little to no role within the epistemology of humanism which, by contrast, remains committed to the study of the "possible" and of what *can* and *cannot* be human. If anything, humanism requires a science of the improbable actually, of the atypical or the abnormal since—"unlikely" as it may be—this improbable remains nonetheless a possible manifestation of what *can* be human. However, the improbable can also only take humanism so far, for it cannot offer a definite characterization of what can *never* be human and, as such, remains fundamentally distinct from the impossible that interests humanism in the end.

It is within this epistemological dispute that we should replace Chomsky's historical opposition to the use of probabilities in linguistics, I believe, as when he tells us that:

> I think that we are forced to conclude [...] that probabilistic models give no particular insight into some of the basic problems of syntactic structure.
>
> Chomsky (1957/2002, p.17)

Or that:

> [I]t must be recognized that the notion of 'probability of a sentence' is an entirely useless one, under any known interpretation of this term.
>
> Chomsky (1968, p.57)

For the argument is not—contrary to what some computational linguistics have interpreted (e.g. Pereira, 2000; Norvig, 2012)—a mere technical point regarding the capabilities of statistical language models to capture nuances of grammaticality—such as the difference between the grammaticality of the sentence *colorless green ideas sleep furiously* and the agrammaticality of *furiously sleep ideals green colorless*. No, what is at stake is a central epistemological question regarding the role that the probable is supposed to play within the language sciences. As far as the humanist perspective in linguistics is concerned, the matter is clear: the critical concept is that of the possible, not the probable. Moro (2016) summarizes the point in his book on *Impossible Languages* when he tells us that:

> To define the class of possible human languages: this is the ultimate aim of linguistics.
>
> Moro (2016, p.1)

And goes on to explain that:

> If we were biologists, we would not hesitate to claim that there are impossible animals: an animal that produces more energy than it absorbs, for example, or an animal capable of indefinite growth. We could make such a claim because all organisms are constrained by physical laws, like entropy or gravity, and it would be relatively easy to formulate the notion of "impossible organisms" to yield a compact description of the animal kingdom (the literature is vast, ranging from [(Thompson, 1917/1992)] to [(Edelman, 1988)] and [(Wesson, 1991)], among others).
>
> Moro (2016, pp.2–3)

Narrow down the concept of "organism" to that of "human being" and you get, I would argue, the perspective put forth by the epistemology of humanism: a commitment to determining what is *possibly* and *impossibly* human.[81]

---

[81] As far as linguistic humanism is concerned, such "laws" would correspond to biological constraints constitutive of the faculty of language and human cognition at large, which would condition which languages could be acquired or learned (see Musso et al., 2003, for a relevant reference on the matter). Note then, that it is only natural to find this critique carry out to modern language models such as ChatGPT as well. In a recent *New York Times* opinion piece, Chomsky (2023) writes indeed that:

> ChatGPT and similar programs are, by design, unlimited in what they can "learn" (which is to say, memorize); they are incapable of distinguishing the possible from the impossible. Unlike humans, for example, who are endowed with a universal grammar that limits the languages we can learn to those with a certain kind of almost mathematical elegance, these programs learn humanly possible and humanly impossible languages with equal facility [(Mitchell & Bowers, 2020)]. (Chomsky, 2023)

Moro et al. (2023) make the matter clear as well when they say that:

> [T]he distinction between possible versus impossible languages cannot be formulated by definition for LLM. (Moro et al., 2023, p.84)

And Bolhuis et al. (2024) emphasize that:

> LLMs can produce 'impossible' languages, not generated by the principles governing all known human languages, just as well as (if not better than) natural language output, and cannot distinguish between them [(Moro et al., 2023)]. (Bolhuis et al., 2024, p.489)

Back to artificial intelligence now: Turing made clear what constitutes his ultimate scientific goal. In a letter to the cyberneticist W. Ross Ashby—probably written around 1946 (see footnote 57 in Turing, 1947/2004, p.375)—he says:

> I am more interested in the possibility of producing models of the action of the brain than in the practical applications to computing.
>
> Turing (1947/2004, p.374)

In his own commentary to Epstein et al.'s (2009) edition of Turing's (1950) original *Mind* paper, Harnad explains that Turing's approach (and so artificial intelligence in general) should be understood as a "branch of reverse bioengineering"—and a branch of the cognitive sciences in particular:

> Another unfortunate terminological choice: "Game" implies caprice or trickery, whereas Turing in fact means serious empirical business. The game is science, the future science of cognition—actually a branch of reverse bioengineering. "Imitation" has connotations of fakery or deception too, whereas what Turing will be proposing is a rigorous empirical methodology for testing theories of human cognitive performance capacity (and thereby also theories of the thinking that presumably engenders it). Calling this an "imitation game" (instead of a methodology for reverse-engineering human cognitive performance capacity) has invited generations of needless misunderstandings [(Harnad, 1992)].
>
> Harnad in (Turing, 1950/2009, p.24)

But a cognitive science of *what*, exactly? Of *human being*, or of *normal being*? Does Turing's approach commits to a science of the *human mind*, or of the *normal mind*? That is the fundamental question I will focus on more specifically in Section 3.3.3.

3.3.2 Normal language

What is normal language? In its simplest and most general formulation, I will say that it is a normative ideal of correct linguistic behavior associated with members of a particular socio-linguistic community (Definition 33).

> **Definition 33 (Normal language)**
> Normal language is a normative ideal of correct linguistic behavior associated with members of a particular socio-linguistic community.

My purpose in this section is now to develop the two core arguments I previously introduced in Section 3.2.1.3, namely, that *language models speak normal language* and yet *nobody speaks normal language*.

Given the previous considerations of Section 3.3.1, part of the first argument seems plain: the fact that language models model "correct linguistic behavior" derives, in a sense, from the straightforward application of the *correctness principle* to the domain of *language*.

If machines are always designed to do something in a predetermined correct way indeed, then it is only natural to expect them to model a "correct linguistic behavior" of some sort when that "something" corresponds to linguistic behavior.

The problem with that argument, however, is twofold. First of all, nothing about the correctness principle actually tells us that this correct linguistic behavior should correspond to the "normative ideal" of a "particular socio-linguistic community", as Definition 33 suggests. Second, the presence of the correctness principle itself is not so obvious in the architectural pipeline of language models. In the textbook definition of language models, for instance—such as the one provided by Jurafsky & Martin (2022) in Definition 34—nothing actually tells us that language models must abide to a correctness criterion of some sort as far as linguistic behavior is concerned.

Definition 34 (Language model)
A language model is a statistical model that assigns probabilities to sequences of words.

Adapted from Jurafsky & Martin (2022, p.31)

The standard protocol for the intrinsic evaluation of language models—usually based on the "perplexity" metric of Definition 35—is no more explicit on the matter either, for all it basically tells us is that a "good" ("better") language model should assign a high(er) probability (in comparison to another model) to a sample of text it has never seen before (see Jurafsky & Martin, 2022, §3.2, for details).

Definition 35 (Perplexity)
The perplexity PP of a language model on a test set is the inverse probability of the test set, normalized by the number of words.
For a test set $W = w_1 w_2 \dots w_N$:

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}}$$

Adapted from Jurafsky & Martin (2022, p.37)

So what is it that makes me say here that *language models speak normal language* and yet *nobody speaks normal language*?

To make sense of those two claims, we must first acknowledge that the very concept of "language model" presupposes a certain conception of what "language" *is*—a very idea of what it is that language models are supposed to be models *of*. For this is what "normal language" is all about: it is a characterization of the very *ontology* of language on which language models rest. The problem with that ontology of language, however, is that it is barely made explicit in practice. So what I propose to do in this section is to make the implicit explicit—to carry out an "archeology" of language models so as to uncover the exact set of assumptions and presuppositions on which they rest and, in doing so, to demonstrate how they must actually conform to a normative ideal of (correct) linguistic behavior, specific to a particular socio-linguistic community, that no individual speaker actually conforms to in practice.

The rest of this section is thus organized around what I take to form the six core properties of normal language. I will argue that normal language is *social* (Section 3.3.2.1), *normative* (Section 3.3.2.2), *objective* (Section 3.3.2.3), *neutral* (Section 3.3.2.4), *single* (Section 3.3.2.5) and *shared* (Section 3.3.2.6).

In Section 3.3.2.1, I will argue that normal language is *social* in the sense that it corresponds to the language of a *community* rather than a single *individual*, and as such constitutes an *idealization* abstracting away from the potential idiosyncrasies of individual speakers—which is why nobody ever speaks *exactly* normal language in the first place.

In Section 3.3.2.2, I will argue that normal language is *normative* in the sense that it characterizes a norm of correctness pertaining to linguistic behavior that is a specification of how speakers *ought* to behave rather than how they *do* behave in practice and which, as such, constitutes an *ideal* from which every individual speaker always deviates—which is why nobody ever speaks normal language in general.

In Section 3.3.2.3, I will argue that normal language is *objective* in the sense that it purports to characterize an objective linguistic reality lying behind its subjective appreciation by each and every individual speaker—which is precisely how normal language rationalizes its characterizing a language that no individual speaker ever speaks in practice.

In Section 3.3.2.4, I will argue that normal language is *neutral* in the sense that the objective linguistic reality it purports to characterize also embodies a (false) promise of neutrality which, in the context of language models, proves attractive to the extent that it promises to guard against unwarranted social *bias*.

In Section 3.3.2.5, I will argue that normal language is *single* in the sense that it rests on the (unfounded) assumption that there exists such a thing as a *unique* normal language, but that any concept of social language always necessarily points at a multiplicity of incommensurable socio-linguistic experiences, so that settling for *one* particular language always ends up settling for the idealized experience of a *particular* socio-linguistic community.

In Section 3.3.2.6, I will argue that normal language is *shared* in the sense that it is always presumed to be common to all members of the corresponding socio-linguistic community, but that what is actually shared is an *ideal* rather than a *practice*—which is precisely what makes the illusion of *shared language* and *shared meaning* hold together.

3.3.2.1 Normal language as social language

When I introduced language models such as GPT-2 (Radford et al., 2019) or GPT-3 (Brown et al., 2020) in Section 3.2.1.2, I said that they could be characterized as statistical (in this case, machine learning) models of language trained on a representative corpus of language use so as to be able to predict the next token in a sequence of strings. I also said that the prediction task itself could be traced back to (Shannon, 1951) and his "guess the next English letter" game, so we might as well start there our investigation into the ontology of language underlying language models.

Shannon (1951) titles his paper "Prediction and Entropy of Printed English" and in a sense he gives us a precious indication: whatever it is that language models are supposed to be models *of*, "English" is one of its concrete and possible extension. But what exactly *is* "English"? Where does it start, and where does it end? Its contours are never quite so clear indeed, for as Chomsky (1993) tells us:

> What are called "languages" or "dialects" in ordinary usage are complex amalgams determined by colors on maps, oceans, political institutions, and so on, with obscure normative–teleological aspects.
>
> Chomsky (1993, p.19)

Normal language thus stems from an everyday concept of language that is notoriously *vague*, which is actually what—for Chomsky (1993) at least—makes it particularly unfit for proper scientific inquiry (see also Chomsky, 2000, especially ch.3).[82] But not for others. Scholz et al. (2022), for instance, insist that:

> Failing to have precise individuation conditions is surely not a sufficient reason to deny that an entity can be studied scientifically. 'Language' as a count noun in the extensional and socio-historical sense is vague, but this need not be any greater obstacle to theorizing about them than is the vagueness of other terms for historical entities without clear individuation conditions, like 'species' and 'individual organism' in biology.
>
> Scholz et al. (2022, p.35)

Be that as it may, in most ordinary use cases, this vagueness is not really an issue. Talking about "language" (or "English") in the general sense of the term—vague or underdetermined as it may be—usually proves good enough for whatever purpose the concept is being put to use. But for us here this underdetermination will not do, precisely because our aim is to understand exactly what it is that a "language model of English" ends up being a model of. This everyday concept of "language" thus has to be refined, and in the process it becomes clear that this refinement cannot be but an *idealization*, for it must necessarily abstract away from the inherent diversity of individual speakers, even across members of the same socio-linguistic community. Santana (2016) makes that point clear in the context of theoretical linguistics, which faces identical challenges when confronted with the same everyday concept of "language":

> [A] linguist can't just adopt the everyday concept of language, and must refine it into a technical notion [. . .] This refinement, however, necessarily involves idealization away from the facts of linguistic diversity. Even a refined technical concept of 'English' will need to elide many of the idiosyncratic differences

---

[82] Chomsky (1986, §2.2 and §2.3) ultimately argues for distinguishing what he calls "E-language"—"externalized" (social) language such as "French" or "English" as per common usage—from "I-language"—the "internalized" language in the mind of a speaker (see also Barber & Garcia Ramirez, 2021, §3, for an overview of the distinction).

> between different speakers of English. So far, no problem. All science idealizes [. . .]
>
> Santana (2016, p.510)

Yet, there remains a fundamental tension behind this idealization, which corresponds to the inherent tension that exists between the *individual* and the *collective* side of every linguistic phenomenon. This tension is at the core of Saussure's (1916/2011) seminal distinction between "language" and (human) "speech" (*langue* and *langage*, in French), which he details as follows:[83]

> But what is language [*langue*]? It is not to be confused with human speech [*langage*], of which it is only a definite part, though certainly an essential one. It is both a social product of the faculty of speech and a collection of necessary conventions that have been adopted by a social body to permit individuals to exercise that faculty. Taken as a whole, speech is many-sided and heterogeneous; straddling several areas simultaneously—physical, physiological, and psychological—it belongs both to the individual and to society; we cannot put it into any category of human facts, for we cannot discover its unity.
>
> Saussure (1916/2011, p.9)

As Saussure makes clear:

> Speech has both an individual and a social side, and we cannot conceive of one without the other.

---

[83] I rely on the French-to-English translation of (Saussure, 1916/2011) by Wade Baskin, who translates *langue* as "language", *langage* as "speech" and *parole* as "speaking". Note, however, that other translations of Saussure's *Course of General Linguistics* may sometimes translate *parole* (and not *langage*) as "speech". Here are some useful terminological clarifications by the editors Meisel & Saussy (2011):

> "language [*langue*]," "human speech [*langage*]" here Baskin first confronts a terminological problem that no English-language translator has satisfactorily resolved. Saussure uses *langage* in the sense of "the human faculty of communication": it is the broadest term in the set. *Langue*, as defined later, is the system of norms accepted and used by members of a speech community (what would ordinarily be referred to as "a language"). *Parole*, which makes its first appearance on p.13, is the act of linguistic expression as performed by an individual trained in (some version of) those norms; for this Baskin often uses "speaking." Saussure is by no means as consistent as he might have been, either in the [*Course of General Linguistics*] or in his handwritten notes, nor does the translator uniformly follow Saussure's terminology. (Meisel & Saussy, 2011, pp.235–236)

In the original French edition of (Saussure, 1916/1971), the corresponding passage reads as follows:

> Mais qu'est-ce que la langue? Pour nous elle ne se confond pas avec le langage ; elle n'en est qu'une partie déterminée, essentielle, il est vrai. C'est à la fois un produit social de la faculté du langage et un ensemble de conventions nécessaires, adoptées par le corps social pour permettre l'exercice de cette faculté chez les individus. Pris dans son tout, le langage est multiforme et hétéroclite ; à cheval sur plusieurs domaines, à la fois physique, physiologique et psychique, il appartient encore au domaine individuel et au domaine social ; il ne se laisse classer dans aucune catégorie des faits humains, parce qu'on ne sait comment dégager son unité. (Saussure, 1916/1971, p.25)

Saussure (1916/2011, p.8)

But "language", on the other hand, is only:

> [...] the social side of speech, outside the individual who can never create nor modify it by himself; it exists only by virtue of a sort of contract signed by the members of a community.
>
> Saussure (1916/2011, p.14)

The point being that, for him, "language" is inherently (if not exclusively) *social*:[84]

> Contrary to all appearances, language never exists apart from the social fact, for it is a semiological phenomenon. Its social nature is one of its inner characteristics.
>
> Saussure (1916/2011, p.77)

From such considerations, two specific conclusions can be drawn for our present purpose. First of all, that if nobody ever speaks *exactly* normal language, it is primarily because normal language always remains but a necessary *idealization* abstracting away from the idiolectal variations of individual speakers. Second, that if normal language does constitute such an idealization, it is because it fundamentally targets the language of a *community* and not that of a particular *individual*. Indeed, when we design a "language model of English", we do not aim to model the linguistic behavior of any particular English speaker, what we aim at is to model the linguistic behavior of the so-called "English speaking community"—even if that community is always necessarily idealized to some extent. To use a convenient analogy, I will say that normal language targets the content of the standard dictionary and not that of a particular subjective lexicon. For if Rodd (2020) tells us that:

> Each person's lexicon is uniquely structured by specific, idiosyncratic linguistic experiences.
>
> Rodd (2020, p.1)

Itkonen (2008) insists on the fact that:

> It goes without saying (or so it seems) that e.g. a dictionary of English is about something that is *common to* or *shared by all* speakers of English, and whatever has these characteristics must be social by definition.
>
> Itkonen (2008, p.279)

---

[84] For an alternative account of the social ontology of language, see, e.g. (Santana, 2016, §4).

And indeed, it is precisely what is *shared by all* rather than *idiosyncratic to some* that normal language intends to capture.

Focusing on the language of a community thus means that normal language must be approached primarily as a *social* entity. It is a practice characterized by a set of rules, norms or conventions purportedly shared by all members of the same (socio-linguistic) community, and it is this social entity that language models primarily intend to model. In sum, normal language is first and foremost *social* language.

3.3.2.2 Normal language as normative language

In his *Course of General Linguistics*, Saussure (1916/2011) repeatedly refers to language as a "social fact", using a terminology then reminiscent of that of the French sociologist Émile Durkheim (see Holdcroft, 1991, §7.3, for a detailed dicussion on the theoretical affinities between Saussure and Durkheim). In the *Chapter II* of his *Introduction* dedicated to the "subject matter and scope of linguistics" and "its relations with other sciences", he explicitly tells us that:

> [L]anguage is a social fact.
>
> Saussure (1916/2011, p.6)

Later on in the *Chapter II* of his *Part One* dedicated to the "immutability and mutability of the sign", he emphasizes once again that:

> [L]anguage never exists apart from the social fact [...]
>
> Saussure (1916/2011, p.77)

But what exactly *is* a "social fact"? For Durkheim (1895/2013):

> A social fact is any way of acting, whether fixed or not, capable of exerting over the individual an external constraint [...] which is general over the whole of a given society whilst having an existence of its own, independent of its individual manifestations.
>
> Durkheim (1895/2013, p.27)

There are, in fact, clear theoretical affinities between Saussure and Durkheim here, since Saussure too considers language—as a *social fact*—to be a supra-individual entity having an existence of its own independent from its particular manifestation in individual speakers. For Saussure indeed, language is this abstract system of (social) *rules*—"necessary conventions", he tells us in his above quote (Saussure, 1916/2011, p.9)—that enables "speaking" (*parole*, in French) and the production of potentially unbounded acts of speech. However, for Durkheim, the "rules" governing social behavior are more than just "conventions", insofar as they are *normative* entities invested in some form of authority conferring them "coercive power" over human behavior (see Durkheim, 1895/2013, p.21). In other words, they are *norms* rather than mere *conventions*.

"Norms", "conventions", is there really a fundamental difference between those two notions? The answer to that question depends, at least in part, on whether one considers conventions *themselves* to be normative entities—which remains a much debated topic. When I introduced the notion of "convention" in Section 2.2.4, I relied on the characterization of (Lewis, 1969/2002) which defines conventions as self-perpetuating behavioral regularities constituting solutions to coordination problems. But this characterization has precisely been criticized for neglecting the potential normativity of conventions. Gilbert (1989, §VI.5), for instance, argues that conventions cannot be reduced to such overt behavioral regularities, for they are not always specifications of how people *do* behave but often, and more fundamentally, specifications of how they *ought* to behave—and as such not always regularities *de facto* but also very much *de jure* (see Rescorla, 2019, §4.1, for details). What is more—and contrary to what Lewis (1969/2002) originally suggested—conventions may not only persist insofar as they constitute equilibria and thus stable solutions to coordination problems—that is, insofar as they are in the best interest of agents who have no reason to depart from them—but also because they crucially have the power to *sanction* any attempt at deviation (see Rescorla, 2019, pp.19–20) and so because they are, in fact, "social facts" in the sense of Durkheim (1895/2013).

When I have used the term "normative" so far in this thesis, it has been mostly in reference to a certain (objective) "requirement of correctness".[85] And indeed, as Winch (1958/2008) tells us, the two notions of "correctness" and "rule-following" are intrinsically connected for:

> [T]he test of whether a man's actions are the application of a rule is [. . .] whether it makes sense to distinguish between a right and a wrong way of doing things in connection with what he does.
>
> Winch (1958/2008, pp.54–55)

Therefore, if we accept the premises that language is a matter of rule-following, we should be able to settle on the normativity of language even before having to settle on the nature of those rules governing linguistic behavior—and in particular, on whether they are *social* or *biological* rules. In other words, so long as we consider that it makes sense to talk about "correct" or "incorrect" linguistic behavior, we should be able to characterize language as being normative. That is, in fact, precisely how I propose to approach the normativity of normal language itself: I will say that normal language is *normative* to the extent that it characterizes a *norm of correctness* with respect to linguistic behavior. Normal language is "correct linguistic behavior" first and foremost, which derives from the straightforward application of the correctness principle to the domain of language.

---

[85] Remember how I introduced my "non-normative subjectivism" in Section 2.2.1 as an approach that no longer required us to objectivize communication success by framing it in so-called "normative terms", and how by that I meant that we no longer had to talk about "correct" or "incorrect" interpretation.

In linguistics and the philosophy of language at large, the question of the normativity of language is usually approached through Wittgenstein's (1953/2009) "private language argument" (see Itkonen, 2008, §2, for an explicit case)—an argument which Biletzki & Matar (2020) nicely summarize as follows:

> [F]or an utterance to be meaningful it must be possible in principle to subject it to public standards and criteria of correctness. For this reason, a private-language [. . .] is not a genuine, meaningful, rule-governed language.
>
> Biletzki & Matar (2020, p.20)

For us here, Wittgenstein's argument proves crucial to the extent that it allows us to completely separate the question of the normativity of normal language from that of its being social. Indeed, when Wittgenstein (1953/2009) tells us for instance that:

> That's why 'following a rule' is a practice. And to *think* one is following a rule is not to follow a rule. And that's why it's not possible to follow a rule 'privately'; otherwise, thinking one was following a rule would be the same thing as following it.
>
> Wittgenstein (1953/2009, §202, pp.87^{e}–88^{e})

Beaney (2009) is careful to explain to us that:

> By 'practice' Wittgenstein means 'practice as a fundamental human activity' rather than 'practice of a community' (though community practices may well constitute the vast majority of such activities); and by 'privately' he does not mean 'in isolation from other people' but something more like 'by reference to one's own interpretation'. [. . .] What he is criticizing is the idea that obeying a rule involves some essential process of 'interpreting' the rule—in the sense of generating or consulting some 'mental representation' of the rule. What counts as following a certain rule is governed by what we do, not by what we think.
>
> Beaney (2009, p.55)

His explanation comes to nuance alternative positions such as that of Dummett (1986), for instance, telling us that:

> A language is a practice in which people engage. There can be solitary practices, in the sense of those in which one engages on one's own; but a practice is essentially social, in the different sense that it is learned from others and is constituted by rules which it is part of social custom to follow.
>
> Dummett (1986, p.473)

Wittgenstein (1953/2009) thus treats the normativity of language independently from the question of its being social, insofar as to him the norms governing linguistic behavior are not intrinsically social in nature.[86] Normal language adopts a similar perspective: the "norm of correctness" that it characterizes does not have to be understood as being exclusively social, for the normativity of normal language does not actually derive from its social nature—and in particular, from a commitment to the intrinsic normativity of the social—but from the correctness principle itself.

Now, what makes it so that *nobody speaks normal language* is that normal language actually characterizes a *normative ideal* of linguistic behavior and constitutes a specification of how speakers *ought* to behave rather han how they *do* behave indeed. Normal language characterizes the behavior of an *ideal* speaker, one who would have mastered to perfection the norm of correctness it characterizes—an ideal speaker with *perfect spelling* and a *dictionary-size* lexicon, for instance.

What is interesting to note indeed about the example of ChatGPT from Section 3.2.1.1 is that it actually contains *no* spelling mistake—which is all the more impressive than it only takes ChatGPT a couple of seconds to generate the requested output. In comparison, I would expect a normal human being, put under similar conditions, to produce at least a few spelling mistakes or typos.[87] Note that I am not saying here that ChatGPT could never be made so as to produce spelling mistakes—or that it would never make any in practice—just that its default behavior is precisely to be designed so as not to make any. And why would it? Why would anyone design a language model that deliberately makes spelling mistakes indeed when that can be easily avoided? Especially when you know that spelling mistakes can cost you millions in online sales or advertising revenue (Everard & Galletta, 2005; Coughlan, 2011; Mozafari et al., 2019), affect your online credibility (Stiff, 2012; Cooper et al., 2020), the perception of the quality of your writing (Kreiner et al., 2002; Liu, 2004; Morris et al., 2012; Jeong et al., 2017), or that a single mistake can even earn you a million pounds lawsuit in the UK (BBC, 2017)? When provided with a clear norm of correctness with respect to linguistic behavior—as with the case of spelling—we can only observe that ChatGPT is made so as to perfectly abide to it. Given the correctness principle, that makes perfect sense: ChatGPT is just designed to follow what is "correct". Yet, this is also precisely where its behavior departs from that of a normal human. Even the best speller, I would argue, is bound to make a spelling mistake at some point—whether it is because of temporary fatigue, distraction, memory lapse, etc., and even though they may otherwise know perfectly well the correct form of the expression at hand.[88] As such, perfect spelling

---

[86] Keucheyan (2010) distinguishes what he calls the "cognitive norms" of Wittgenstein from the "social norms" of Durkheim, and discusses similarities and differences between the two.

[87] An intuition that is widely shared since, as we will see in Section 3.3.3.1, human judges in practical Turing tests typically rely on the presence/absence of spelling mistakes in participants' output to discriminate between human and machine participants.

[88] "Mistakes are a fact of life", Lunsford & Lunsford (2008) tell us in their review of first year college students writing in the USA, where they report a typical error rate of 2 to 3 mistakes per 100 word (see Lunsford & Lunsford, 2008, Table 8). Note that their concept of "error" includes notions such as "vague pronoun reference" or "unnecessary comma" that probably extends much beyond what is relevant to our present discussion, but that at the same time their concept of "spelling error" excludes notions such as

probably constitutes the prototypical example of a normative ideal of linguistic behavior that we all strive to satisfy and yet never quite manage to in practice.

The same argument can then be carried out with respect to the lexicon this time, since ChatGPT's lexical knowledge resembles more what could be found in a standard dictionary than in a subjective human lexicon. Just think of how, for instance, ChatGPT is capable of providing definitions for so-called "rare" English words such as "gobbledygook", "ulotrichous" or "kakorrhaphiophobia"—words that I did not know myself before carrying out this research—definitions that align pretty well with those of the Merriam–Webster online dictionary (see Appendix A.2 for details). Beyond those particular examples and the case of rare words in general, my argument is that *no* human speaker can ever be expected to know *all* the words in the dictionary anyway (remember Brysbaert et al., 2016, in footnote 67 indeed) and yet we have no reason to expect that ChatGPT would be deliberately made so as to have partial lexical knowledge if that can be avoided.[89]

I said in the previous Section 3.3.2.1 that normal language does not correspond to the language of any individual speaker, but what we see here is that normal language does not correspond to the language of a "normal" speaker (typical or average) either. In the case of spelling and the lexicon, the point is clear: normal language does not correspond to the normal/average linguistic behavior of a normal/average speaker made of average spelling and average lexical knowledge. It corresponds to the *exceptional* linguistic behavior of an *ideal* speaker made of *perfect* spelling and *full* dictionary-size lexical knowledge. In the end, nobody speaks normal language but not only because normal language is an *idealization*—and abstraction from which most speakers should remain pretty close to nonetheless. Nobody speaks normal language because normal language is also (and more fundamentally) an *ideal*—a normative ideal of exceptionality, in fact, from which most people are actually *far off*.[90] All this because normal language is *normative* language.

3.3.2.3 Normal language as objective language

Treating normal language as *social* language as I did in Section 3.3.2.1 necessarily implies having to commit to a particular *social ontology*, and there are in fact many good reasons to mobilize the theoretical apparatus of Durkheim for that purpose. Durkheim's "social realism" (Definition 36) proposes indeed to treat social facts as *data*, which makes it possible to objectivize social facts by leveraging the full power of the metrological sciences and, in our case, to make use of *statistics* in particular to objectivize language itself.

"missing word/comma/hyphen/capitalization" that I would definitely include in our concept of "spelling mistake" (see Lunsford & Lunsford, 2008, and notably Table 4 and Table 7, for details). Else, Flor et al. (2015) report, in a different context, that only 10.7% of essays from native speakers of English in the TOEFL or the GRE exams contain no spelling mistake (see Flor et al., 2015, Table 3) and that this is once again in conditions that are in many ways more favorable than that of ChatGPT in the example above, which produced its output in a single pass and within a couple of seconds only.

[89] Once again, my point here is not to argue that ChatGPT already *does* display full lexical knowledge in every possible context, but only that if it could be made so it would, given both the correctness principle and the tasks it is usually being put to use.

[90] Remember that it is a Galtonian ideal of exceptionality located at the far right of the cumulative normal curve in Figure 3.2, precisely *far away* from the statistically normal/average point of reference.

> **Definition 36 (Social realism)**
> Social realism is an epistemological position which posits that social facts can be treated as objects of potential knowledge—that is, as *data*—both external to the observer—given, offered, indeed "imposing" themselves on observation—and external to individuals themselves, who are always surrounded, preceded and outlasted by them.
>
> Adapted (extensively) from Lukes (1982/2013, p.xxi)

In *The Rules of Sociological Method*, Durkheim (1895/2013) begins his chapter on the *Rules for the Observation of Social Facts* by stating his foundational methodological principle:

> The first and most basic rule is *to consider social facts as things.*
>
> Durkheim (1895/2013, p.29)

It is easy to understand how such a principle could be attractive epistemologically speaking: if social facts are "things" indeed then they can be *measured* and so "objectivized" in some sense. It is, in fact, this very promise of objectivization that underlie the "metrological realism" of statistics in the nineteenth century, as Desrosières (2001) explains:

> *Metrological realism* derives from the theory of measurement in the natural sciences that is complemented, in the social sciences, by the sampling method. The object to be measured is just as real as a physical object, such as the height of a mountain. The vocabulary used is that of reliability: accuracy, precision, bias, measurement error (which may be broken down into sampling error and observation error), the law of large numbers, average, standard deviation, and estimation by the least-squares method [(Stigler, 1986; Hacking, 1990)]. This terminology and methodology was developed by eighteenth-century astronomers and mathematicians, notably Gauss, Laplace, and Legendre. The core assumption is the existence of a reality that may be invisible but is permanent—even if its measurement varies over time. Above all, this reality is independent of the observation apparatus. In a sense, this is the dream of the statistician and the specialist in quantitative social sciences: the possibility of making the metrology of these sciences equivalent to the proven methodologies of the natural sciences. This may be seen as a benchmark, an ideal to which statisticians aspire, despite an awareness that their objects do not display all the properties assumed by the methodology. We could describe this as the lost paradise of the social sciences, which would have liked to have been endowed with the same persuasiveness as the natural sciences of the nineteenth century.
>
> Desrosières (2001, pp.340–341)

The *social realism* of Durkheim thus interfaces well with a certain "statistical realism" (Definition 37) which can itself apply equally well to both social and natural phenomena. It is therefore not surprising to see the two fields of *sociology* and *statistics* having developed somewhat concurrently—sociology finding in the quantitative methods provided by statistics one way to ground its own objectivity (see Schutt, 2012/2020, for an overview).[91]

> **Definition 37 (Statistical realism)**
> Statistical realism is an epistemological position which posits that we can access reality through statistical aggregates or statistical inference based on data.
>
> Adapted from Benbouzid (2023, p.37)

Those epistemological considerations are still very much relevant for us today, for as Benbouzid (2023) tells us:

> [W]ith the culture of machine learning [...] statistical realism is present once again, but in a new form, with the argument that a permanent, temporal and spatial expansion of increasingly granular, algorithmically processed data brings the model ever closer to reality [...]
>
> Benbouzid (2023, p.37)

The methods may change but the presuppositions remain the same, and notably that "data science" as a whole constitutes an epistemology that can give us access to an objective reality that is otherwise inaccessible to our subjective experience. As McQuillan (2018) nicely summarizes indeed:

> Data science is not simply a method but an organising idea. [...] It appears to reveal a hidden mathematical order in the world that is superior to our direct experience.
>
> McQuillan (2018, p.253)

The underlying epistemology thus rests on a comparative opposition between the *objective* and the *subjective*, and on the assumption that there exists an objective reality that is distinct from our subjective perception of it—and, in the case of social facts, from its singular manifestation in each individual at a given point in time. But the idea is also that this objective reality is in fact *better*—"superior" as McQuillan (2018) tells us—and so in a sense *truer* than our own subjectivity, which justifies its becoming the focus of scientific inquiry. Durkheim (1895/2013) makes that point clear in his *Rules for the Observation of Social Facts* when he tells us that:

---

[91] Durkheim himself made direct use of Quetelet's (1835/2013) statistically-based "social physics" in *The Division of Labor in Society* (Durkheim, 1893/1984)—and his statistical construct of the "average man" in particular—before distancing himself from it later on in *Suicide* (Durkheim, 1897/2005). See (Desrosières, 1993/1998, pp.96–101) for details.

> [I]t is [the objective] rules and not the cursory view we have of them which constitute the subject matter of science, just as the subject matter of physics consists of actual physical bodies and not the idea that ordinary people have of it.
>
> Durkheim (1895/2013, p.34)

In linguistics, this opposition between (objective) reality and (subjective) perception is what we sometimes find behind the grammatical/acceptable dichotomy—or the opposition between what *is* correct and what *seems* correct (see Itkonen, 1978, §5.2, for details). The normative aspect of the opposition between objective and subjective itself is also salient in (Wittgenstein, 1953/2009) once again, which explicitly relegates the idea of (linguistic) subjectivity and subjective understanding to the realm of mistakes and errors:

> Let us remember that there are certain criteria in a man's behaviour for his not understanding a word: that it means nothing to him, that he can do nothing with it. And criteria for his 'thinking he understands', attaching some meaning to the word, but not the right one. And lastly, criteria for his understanding the word correctly. In the second case, one might speak of a subjective understanding. And sounds which no one else understands but which I '*appear to understand*' might be called a "private language".
>
> Wittgenstein (1953/2009, §269, p.101^{e})

Wittgenstein (1953/2009) insists throughout his private language argument that "correctness" in language ought to be characterized *objectively*, since:

> [If] whatever is going to seem correct to me is correct [. . .] that only means that here we can't talk about 'correct'.
>
> Wittgenstein (1953/2009, §262, p.99^{e})

This "objectivity" necessarily entails something *external* to individuals themselves, since:

> [J]ustification consists in appealing to an independent authority [. . .]
>
> Wittgenstein (1953/2009, §265, p.100^{e})

At its core, the objectivization of language thus fundamentally rests on the *qualitative* marginalization of subjectivity, since its presence and potential pervasiveness is usually acknowledged but nonetheless relegated outside of the scope of inquiry—subjectivity does not matter indeed, for the proper object of study is *language*, the objective linguistic reality that lies *behind* our subjective perception and its potential singular manifestation in each and every one of us. And as Harris (1954) tells us:

> [T]he structure of language does not necessarily conform to the structure of subjective experience, of the subjective world of meanings.

Harris (1954, p.151)

As far as normal language is concerned, those considerations prove crucial insofar as they accommodate if not actually rationalize the fact that *nobody speaks normal language*. Nobody speaks normal language indeed but that is all fine because normal language precisely intends to characterize this objective linguistic reality that lies *behind* our deviant subjectivities—an objective reality that now becomes accessible thanks to the power of statistical objectivization. Nobody speaks normal language but that is actually expected since our respective idiolects are always "mistaken" one way or another—always partial and partially erroneous grasp of this objective, correct and true linguistic reality anyway. Nobody speaks normal language but that is all fine in the end because, in a word, normal language is actually *objective* language.

3.3.2.4 Normal language as neutral language

Nothing probably better illustrates this commitment to objectivity in computational linguistics than the use of the term "debiasing" by the language modeling literature suggesting, even if only implicitly, that there could exist an "*un*-biased"—potentially *true*, *objective* and *neutral*—state of language, if not of the (social) world entirely (e.g. Bolukbasi et al., 2016; Kaneko & Bollegala, 2021a,b; Zhou et al., 2021; Lauscher et al., 2021; Friedrich et al., 2021).[92] When applied to the generation of vector-based representations of lexical meaning in particular, statistical models of language have been shown to exhibit various sorts of *social biases*—racist, sexist, classist, ableist or queerphobic, among others (see Caliskan et al., 2017; Garg et al., 2018; Blodgett et al., 2020; Cao & Daumé III, 2021; Field et al., 2021; Abid et al., 2021; Hassan et al., 2021; Navigli et al., 2023, for reviews)—but the problem actually extends well beyond lexical meaning and all the way to natural language processing and (machine-learning-based) artificial intelligence itself since, as Birhane (2021) rightly notes:[93]

> When [machine learning] systems "pick up" patterns and clusters, this often amounts to identifying historically and socially held norms, conventions, and stereotypes.

Birhane (2021, p.44)

---

[92] Some of the references cited above target "word embeddings" rather than "language models" *per se*, but note that word embeddings can themselves be understood as language models applied to the generation of lexical semantic representations. In traditional word embeddings models such as WORD2VEC, for instance (see Mikolov et al., 2013) the output vectors modeling lexical meaning correspond to the hidden layer of a neural network otherwise trained on the regular task of predicting a target token given a (sequence/set of) context token(s)—which is just here a "looser" version of the traditional language modeling task of predicting the next token in a sequence of strings—or a variant of that task itself consisting in predicting the context token given a target token (see Gastaldi, 2021, §2.2, for details).

[93] Which is actually why, in an early paper on the question of biases, Herbelot et al. (2012) argue that statistical methods could be useful for philosophical inquiry as they would make those "biases" observable and more apparent.

Of course, one could argue here that the use of the term "debiasing" should simply be understood pragmatically and as referring merely to the desire to remove such unwarranted social biases, without necessarily implying that there could exist such a thing as a "true", "objective" and "neutral" state of language or the social world entirely. We could be willing to "debias" our models by removing specific types of social biases, without necessarily implying that the debiased models so produced would actually be truer, more objective, or even *less* biased than they were before.[94] Yet, if you take a closer look at the debiasing literature, you actually find references such as (Kaneko & Bollegala, 2021b), for instance, typically telling us that:

> Word embeddings trained on large corpora have shown to encode high levels of unfair discriminatory gender, racial, religious and ethnic biases. In contrast, human-written dictionaries describe the meanings of words in a concise, objective and an unbiased manner.
>
> Kaneko & Bollegala (2021b, p.212)

So the matter is far from obvious, since Kaneko & Bollegala (2021b) do seem to presume here that dictionaries characterize "objective" and "unbiased" meaning—and so that such a thing as an objective and neutral state of language and meaning exists in the first place.

The example of dictionaries is all the more interesting to us here than the field of lexicography itself has historically faced similar controversies regarding the presence of social bias in its output.[95] Back in 1924 indeed, the *Oxford English Dictionary*, for instance, faced significant backlash for the presence of outright antisemitic exemplar senses such as "unscrupulous usurer or bargainer" or "to cheat, overreach" under the lexical entry for the word *Jew* (see Burchfield, 1989, pp.109–110, for details).[96] Both the head of the *Oxford University Dictionary* R. W. Chapman and the editor of the *Pocket Oxford Dictionary* H. W. Fowler then reacted to the controversy along similar lines, opposing to their detractors the requirement of objectivity of the lexicographer commanding them to document language as it *is* used and not as it *should* be used:

---

[94] One could argue indeed that language and the social world are in fact always intrinsically "biased"—in the sense that we always have to arbitrate between different systems of values—so that the term "debiasing" actually refers to the process of shifting from one biased state to another, given an arbitrary set of values collectively negotiated and agreed upon: we move away from one biased state that we collectively deem "harmful" or "undesirable" and towards another biased state that we deem "desirable" this time, without any claim at the universality, intemporality or immutability of such states, which could in fact always be collectively renegotiated at any point in time.

[95] Critical lexicography today even tells us that:

> Lexicography and its products, dictionaries, are never value-free, apolitical or asocial. Instead, they are subject to ideology, power and politics. (Chen, 2019, p.1)

So the objectivity and neutrality of dictionaries is far from obvious, contrary to what Kaneko & Bollegala (2021b) seem to assume (see Kachru & Kahane, 1995, for an overview).

[96] See also (Benson, 2001) for another reference on the issue. More recently, Chen (2019) mentions a 2016 controversy surrounding the use of the sexist exemplar phrase "a rabid feminist" to illustrate the use of the term *rabid* in *The Oxford Dictionary of English* (see Chen, 2019, pp.1–2).

> [I]t is no part of the duty of a lexicographer to pass judgement on the justice or propriety of current usage.
>
> Robert William Chapman, as cited in (Burchfield, 1989, p.110)

> The dictionary-maker has to record what people say, not what he thinks they can politely say: how will you draw the line between this insult to a nation and such others as 'Dutch courage', 'French leave', 'Punic faith', 'the Huns', 'a nation of shopkeepers', and hundreds more? The real question is not whether a phrase is rude, but whether it is current.
>
> Henry Watson Fowler, as cited in (Burchfield, 1989, p.110)

The conception of objectivity in modern lexicography thus appeals to a certain *descriptivism* that actually opposes the early *prescriptivism* of the field—such as when Samuel Johnson (1747/2021) defended his original *Plan of a Dictionary of the English Language* by arguing that:

> Barbarous or impure words and expressions, may be branded with some note of infamy, as they are carefully to be eradicated wherever they are found; and they occur too frequently even in the best writers.
>
> Johnson (1747/2021, p.393)

Prescriptivism today in lexicography is most certainly disqualified and, as a doctrine, relegated to the realm of unwarranted snobbery and elitism from individuals more eager to reinforce the power and authority of their social group than to contribute anything potentially useful to the language sciences (see Hutton, 2009, pp.20–34 for a discussion). Yet, descriptivism does not signify the end of normativity in lexicography altogether, for as Cameron (1995) rightly emphasizes:

> [Both] prescriptivism *and* anti-prescriptivism invoke certain norms and circulate particular notions about how language ought to work. [. . .] On that level, 'description' and 'prescription' turn out to be aspects of a single (and normative) activity: a struggle to control language by defining its nature.
>
> Cameron (1995, p.8)

In effect, descriptivism still commits to a notion of "correct linguistic behavior", but shifts its specification from the judgment of a few pundits to the practices of the whole community of language users. What is "correct" as far as language is concerned becomes how people talk and so how they use language collectively. In other words, *objectivity is use.*

This conception of "objectivity-as-use" is precisely what we find at the core of the ontology of normal language underlying language models. In fact, I will go as far as to argue that dictionaries themselves constitute the "theoretical ancestors" of language models, in the sense that they both rest on the very same ontology of normal language and are produced by the very same epistemology. Both dictionaries and language models indeed start from the same point: a *corpus*, purportedly representative of the target language at hand. Both commit to a *social* ontology of language which purport to characterize the linguistic behavior of a community rather than that a single or particular individual. Both contain *normative* aspects, since both characterize in their own way what is "correct" as far as linguistic behavior is concerned, and both actually constitute normative *ideals* in the sense that both characterize how people *ought* to behave rather than how they *do* behave in practice. Nobody can be said to know all the words in the dictionary indeed, although we may very much like to. Last, but not least, both employ *statistical means* to objectivize language and characterize the purportedly objective linguistic reality lying behind our respective subjectivities. In the case of language models, those statistical means correspond to the aggregation of statistical patterns of sequences of tokens which distribution forms the core of the prediction model—exploiting what Shannon (1951) refers to in his original paper as the "statistical structure" of language.[97] In the case of dictionaries, those statistical means correspond to the frequency of occurrence of lexical items which determine notably whether a given item is attested frequently enough within the representative corpus at hand to warrant its presence in the dictionary in the first place (or whether it should rather be excluded and disqualified as an unrepresentative idiosyncrasy or an individual "mistake" of some sort).

Is *objectivity-as-use*, then, sufficient to guarantee the neutrality of language models—even if only under the pragmatic interpretation of the term? Everything depends on *whose use* it is that the corpus is actually documenting, since as Rogers (2021) tells us:

> A typical corpus for training language models, or really any [Natural Language Processing] dataset, is likewise a sample of speech of a certain group of people, who have their linguistic preferences and sets of values. Consequently, that sample, whether it is coherent or not, and whether it was collected with any specific intentions, represents a certain "picture of the world". Moreover, the purpose of using this data for training is to create a system that would encode that view of the world and make predictions consistent with it. But a typical [Natural Language Processing] dataset [footnote omitted] currently has few specifications of the demographics, dialects, or the range of domains and linguistic phenomena it covers. Unfortunately, it does not mean that the result is

---

[97] See, for instance, the following passage:

> In a previous paper [(Shannon, 1948)] the entropy and redundancy of a language have been defined. The entropy is a statistical parameter which measures, in a certain sense, how much information is produced on the average for each letter of a text in the language. If the language is translated into binary digits (0 or 1) in the most efficient way, the entropy H is the average number of binary digits required per letter of the original language. The redundancy, on the other hand, measures the amount of constraint imposed on a text in the language due to its **statistical structure**, e.g., in English the high frequency of the letter E, the strong tendency of H to follow T or of U to follow Q. (Shannon, 1951, p.50, my emphasis in bold)

> some abstract "standard" or "neutral" language. It is some kind of interpolation from the mixture of signals in the data that we have very little idea about.
>
> Rogers (2021, p.2187)

But not only. For as Gastaldi & Pellissier (2021) rightly explain, a corpus itself is always the product of an entire process of data curation, normalization and organization, and as such constitutes an entity that is not so much "discovered" than *constructed* by linguists and computational linguistics alike:

> Certainly, corpora are not disembodied devices unrelated to extralinguistic dimensions. Despite a general tendency to treat corpora as neutral and unbiased datasets, it is in the nature of a corpus to be an expression of concrete practices, as well as of a partial way of recording, selecting, normalizing and organizing them.
>
> Gastaldi & Pellissier (2021, p.577)

Therefore, the neutrality of language models—or of dictionaries, in fact—depends as much on *whose use* it is that the underlying corpus is actually documenting than on *how* it is constructed and so *by whom*—considerations that extend much beyond corpus construction itself and all the way to language modeling or dictionary building.[98]

In any case, what those considerations suggest is that objectivity and neutrality are clearly separate matters, and that grounding the objectivity of language models in language use does not suffice to guarantee their being unbiased. Be that as it may, the aspiration to neutrality remains, and as the literature on debiasing critically suggests, the aim of the field is still to build neutral language models exempt from specific social biases. Neutrality is therefore a given, even if it only refers to a specific biased state of language and the world in practice. The objectivity of use must therefore go through the filter of neutrality if it is to produce an unbiased language model so that, in the end, what is modeled is never the language use of *anyone* and *everyone* but always the use of a *particular* socio-cultural community with specific values and specific worldviews. In sum, normal language is *neutral* language.

3.3.2.5 Normal language as single language

Asking *whose use* it is that language models are objectivizing when modeling normal language points at yet another critical assumption of their ontology of language: the fact that this objective linguistic reality that normal language purportedly characterizes corresponds to a *single*, *unique* and *shared* reality between speakers of the same language. Johnstone

---

[98] Many arbitrary decisions have to be carried out with respect to language modeling or dictionary building indeed. Just think of how, for instance, one must settle for a particular threshold value so as to determine what constitutes a "sufficiently attested" lexical item, or how one must settle for a particular aggregation algorithm so as to gather statistical data in a language model. All those arbitrary decisions explain why we may always end up with *different* dictionaries or *different* language models of the "same" language in the end.

(2000) makes that point clear in the context of structuralist linguistics when she tells us that:

> Linguists who study *langue* thus study something that is by definition superindividual, located in individuals only the way a piece of software may be installed on multiple computers. The object of study for structuralist linguistics is a “shared” system, that is a system that is of interest only insofar as it can be treated as identical from individual to individual.
>
> Johnstone (2000, p.408)

Indeed, Saussure (1916/2011) explicitly states that:

> Language exists in the form of a sum of impressions deposited in the brain of each member of a community, almost like a dictionary of which identical copies have been distributed to each individual [...] Language exists in each individual, yet is common to all.
>
> Saussure (1916/2011, p.19)

However, can we really consider that all speakers of the same language share the same norm of correctness *vis-à-vis* their language—the same normative ideal of what it means to speak their language correctly? Is there such a thing as a “single” or “unique” normal language in the first place? Not for Eagleton (1983/2008) at least, who argues that:

> The idea that there is a single ‘normal’ language, a common currency shared equally by all members of society, is an illusion. Any actual language consists of a highly complex range of discourses, differentiated according to class, region, gender, status and so on, which can by no means be neatly unified into a single homogeneous linguistic community. One person’s norm may be another’s deviation [...]
>
> Eagleton (1983/2008, p.4)

Indeed, when we talk about “English”, we potentially refer to many different norms of correctness: some which may command that we spell [ˈɔrgəˌnaɪz] as *organize* rather than *organise*; others that we pronounce *tomato* as [təˈmɑːtəʊ] rather than [təˈmeɪtoʊ]. But what Eagleton (1983/2008) also critically suggests here is that all those distinct norms actually correspond to different socio-linguistic experiences of English—and so to *different norms* of English themselves rather than just different subjective appreciations of the same objective norm of English. There is “British English”, “Australian English” or “African American Vernacular English”, for instance—among many other variants of English for which we may not have clear and definite labels actually—each corresponding to a particular socio-linguistic community and each potentially opposing the others with respect to what it considers to fall under “correct English behavior”. Be that as it may, all those *dialectal* or *sociolectal* variations of “English” remain but mere idealizations in any case, for they must all necessarily abstract away from the *idiolectal* variations of their individual members who all presumably have but "partial" and "partially" erroneous appreciation of those objective dialectal or sociolectal norms. But what if those idiolectal variations did not constitute subjective appreciations of the same objective dialectal or sociolectal norm, but different individual norms themselves?

Now would probably be the right time to ask *why* we all speak different idiolectal variations of English, and to acknowledge that if we never quite speak the same idiolect of English, it is primarily because we never quite have the same socio-linguistic experiences of English at the individual level and because, in the end, we all experience the (social) world in our own way. Johnstone (2000) states the matter quite clearly, I believe, when she tells us that:

> The claim that nothing except innate predilections underlying human linguistic competence could be completely homogeneous from speaker to speaker is not, however, fundamentally an empirical claim. It is rather the logical consequence of the fact that each human's world of experience is different. Knowledge of language is fundamentally private and individual, and it is impossible that two people could do things with language the same way. This is the result of the fact that people are not born knowing how to talk. Although we say that many American children "learn English," in fact no two learn exactly the same thing. One person's language is different from another's because each individual has a different set of linguistic memories and each may make different generalizations on the basis of what he or she hears. This is obvious in the case of children, who have accumulated relatively few linguistic memories, and we regularly notice their idiosyncratic generalizations. But adults' generalizations are idiosyncratic, too, both in the sense that they may in fact be different from anyone else's generalizations and in the more fundamental sense that they are logically independent of anyone else's generalizations. It is often obvious that people have different vocabularies and their sense of the nuances of word meaning varies, but it is equally the case that people have different grammars.
>
> Johnstone (2000, p.411)

The assumption of a single, unique and shared normal language thus appears for what it is: a fundamental marginalization of subjectivity that rests on a "myth of homogeneity" which systematically downplays the singularity of our respective socio-linguistic experiences.

To make sense of that argument, it may be useful here to make a quick detour by the lexicon once again, since Clark's (1998) notion of "communal lexicons" precisely addresses the presuppositions underlying the myth of homogeneity—namely the *quantitative* and *qualitative* marginalization of subjectivity—while at the same time illustrating how the singularity of our respective lexical competences remains fundamentally grounded in the singularity of our own socio-linguistic experiences. Clark (1998) begins with the following crucial consideration regarding our *personal* or *subjective* lexicon:

> You and I each have a *personal lexicon*, a stock of words we know to varying degrees, and even though we both speak English, our personal lexicons differ. The differences may be enormous.
>
> Clark (1998, p.63)

Here, he specifically targets the *quantitative* marginalization of subjectivity which treats the systematic differences between our respective personal lexicons as *marginal* rather than *significant*—or even "enormous", as he himself suggests. He then goes on to argue:

> 'Of course our personal lexicons differ,' you might retort, 'but that is an accident of our life histories. The differences are haphazard and have no bearing on theories of language or language use.' But are these differences truly haphazard? I will argue no. Personal lexicons have systematic differences that we keep close track of. These differences help determine what we mean and what we take others to mean. They play an integral role in the formulation and interpretation of utterances. They bear directly on how language is structured and used.
>
> Clark (1998, p.63)

There, he targets the *qualitative* marginalization of subjectivity this time, which though acknowledging the potential pervasiveness of subjectivity—"of course our personal lexicons differ"—relegates it nonetheless *outside* of the scope of investigation, telling us basically that it does not matter for it has "no bearing on theories of language and language use". Then, Clark introduces his critical notion of "communal lexicon":

> At the core of the argument are what I will call *communal lexicons*. A communal lexicon is the vocabulary associated with a community of people—for example, physicians, football aficionados, San Franciscans—who are distinguished by their common knowledge of a particular field of expertise—medicine, football, San Francisco. We each belong to many communities—I might be a physician, football aficionado, *and* San Franciscan—and know the lexicons of these communities. The claim is, we keep track of our communal lexicons. When I speak to you, I try to select words from the lexicons of the communities I believe you and I both belong to. You and I cannot communicate successfully without distinguishing among our communal lexicons.
>
> Clark (1998, pp.63–64)

Clark crucially reminds us here that we, as individual speakers, are always the product of a great variety of distinct socio-linguistic experiences but that we, as individual speakers, are also the product of *particular* or *singular* combinations of otherwise shared socio-linguistic experiences (although the latter claim is probably more implicit in his account than the former). We are all the product of shared socio-linguistic experiences, but at the same time never quite the same, which is precisely what grounds the singularity of our personal lexicons: one English speaker will be a San Franciscan, a physician and a football aficionado when another will be a New Yorker, a carpenter and a football aficionado; another a Londoner, a carpenter and a soccer aficionado this time; and yet another none of those, but also potentially so much more, etc. The singularity of our (social) experiences at large is what grounds the singularity of our respective subjectivities and ultimately contributes to making us all individually unique.

Does it really make sense, then, to idealize the English speaking community as a homogeneous whole when we probably never quite share the same socio-linguistic experiences of English? What is more, the idealization of normal language only works insofar as it purports to characterize a *reasonable* approximation over all its possible individual variations, but is this approximation really "reasonable" if the systematic differences across our respective idiolects are in fact "enormous" rather than *marginal* as Clark (1998) suggests?

Some may argue here that there could still exist a common denominator between all our idiolects of English—a single set of shared socio-linguistic experiences that would warrant referring to them all as "English" in the first place and could ultimately ground a common normative ideal of (correct) English behavior. To counter that argument, it actually proves useful to return to Wittgenstein (1953/2009) once again to make use of his concepts of "language-game" and "family-resemblance". Asking *whose* use it is that language models are objectivizing with normal language as I did in the beginning of this section suggests that I would be targeting Wittgenstein's *meaning-is-use* doctrine at some point.[99] But that is not actually the case. In fact, I believe that Wittgenstein's theory of language can precisely help us get rid of what is left of the myth of homogeneity and craft the definite argument against the uniqueness of normal language.

To begin, let me start by emphasizing that Wittgenstein (1953/2009) fundamentally encourages us to:

> [...] make a radical break with the idea that language always functions in one way, always serves the same purpose: to convey thoughts—which may be about houses, pains, good and evil, or whatever.
>
> Wittgenstein (1953/2009, §304, p.109^{e})

He insists indeed on the fact that:

> [W]e are so much accustomed to communicating in speech, in conversation, that it looks to us as if the whole point of communicating lay in this: that

---

[99] The doctrine in question refers to that passage of the *Philosophical Investigations* where Wittgenstein tells us that:

> For a *large* class of cases of the employment of the word "meaning"—though not for *all*—this word can be explained in this way: the meaning of a word is its use in the language. Wittgenstein (1953/2009, §43, p.25^{e})

Though it is probably phrased slightly more straightforwardly in his posthumous *Philosophical Grammar*:

> The use of a word in the language is its meaning. Wittgenstein (1974, §23, p.60)

> someone else grasps the sense of my words—which is something mental—that he, as it were, takes it into his own mind. If he then does something further with it as well, that is no part of the immediate purpose of language.
>
> Wittgenstein (1953/2009, §363, p.121^{e})

To him, language is a “game” (§7), a fundamental human *activity* that constitutes what he calls a “form of life” (§23), which is itself a system that gives rise to a shared behavior across humans (§206). His point being that “language-games” are fundamentally *plural*, as they can be as diverse as:

> Giving orders, and acting on them—
> Describing an object by its appearance, or by its measurements—
> Constructing an object from a description (a drawing)—
> Reporting an event—
> Speculating about the event—[. . .]
> Forming and testing a hypothesis—
> Presenting the results of an experiment in tables and diagrams—
> Making up a story; and reading one—
> Acting in a play—
> Singing rounds—
> Guessing riddles—
> Cracking a joke; telling one—
> Solving a problem in applied arithmetic—
> Translating from one language into another—
> Requesting, thanking, cursing, greeting, praying.
>
> Wittgenstein (1953/2009, §23, p.15^{e})

What is more, the definite list of language-games can never be settled once and for all, since all the things that we do with language—all the language-games that we play—always keep changing as:

> [. . .] new types of language, new language-games, as we may say, come into existence, and others become obsolete and get forgotten.
>
> Wittgenstein (1953/2009, §23, pp.14^{e}–15^{e})

Crucial to our considerations, then, is that fact that to Wittgenstein, language-games do not actually share a single common feature together:

> Instead of pointing out something common to all that we call language, I’m saying that these phenomena have no one thing in common in virtue of which we use the same word for all a but there are many different kinds of *affinity* between them. And on account of this affinity, or these affinities, we call them all “languages”.

> Wittgenstein (1953/2009, §65 p.35^{e})

They only share a set of "different kinds of *affinity*" indeed, which constitutes what he calls a "family resemblance":

> I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family—build, features, colour of eyes, gait, temperament, and so on and so forth—overlap and criss-cross in the same way.—And I shall say: 'games' form a family.
>
> Wittgenstein (1953/2009, §67, p.36^{e})

Now, Wittgenstein's notion of human activity—as we saw in Section 3.3.2.2—extends much beyond *social* activities, and his concept of norm, in turn, is probably more *cognitive* than *social*. But even so we could probably stretch his argument here to argue that each and every one of our idiolects is the product of a particular set of socio-linguistic experiences which all individually define their own "game" with their own rules and their own norms. If those *games* all share but a *family resemblance* then, it should not be possible to identify a single common denominator between all our idiolects that would be able to ground a single common normative ideal of (correct) English behavior. Doing so would allow us to explain how the myth of homogeneity can hold together in the first place, as it would be able to account for the fact that we do always share "something" with other speakers of English as far as our idiolects and socio-linguistic experiences are concerned—hence our impression of homogeneity—but that this "something" is in fact never quite the same from one speaker to the next. The myth of homogeneity is but a myth after all.

Back to the uniqueness of normal language now and the question is: which norm will a language model of English eventually objectivize and enforce? Will it spell [ˈɔrɡəˌnaɪz] as *organize* or *organise*? Will it use *soccer* or *football*? Will it say "the government *are*" or "the government *is*"? Objectivizing English as a single, unique normal language necessarily means that one will have to arbitrate among various distinct and potentially conflicting norms of English, and so that any given language model will necessarily correspond to a particular socio-historically and socio-culturally situated version of English—idealized or even "fictitious" as it may be. Arbitrating across various distinct norms of English necessarily entails that normal language will have to arbitrate among various distinct and potentially irreconcilable socio-linguistic experiences of English, in turn—*including* some while *excluding* others—so that it will ultimately end up structuring and organizing the collection of English idiolects with respect to the normative point of reference that it defines. Nobody will speak normal language for sure, but some will probably speak something "closer" to it than others. All this, in fact, because normal language is also *single* language.

3.3.2.6 Normal language as shared language

In detailing the assumption of a shared *langue* underpinning the whole structuralist enterprise, Johnstone (2000) is careful to remind us of its socio-historical context:

> Historically, structuralist linguistics is partly rooted in the nineteenth-century politics. The ideology of nationalism depended on the notion of the autonomous, super-individual language because a nation was defined, in part, linguistically. Arguments in favor of the nation-states that replaced authoritarian feudal systems were often based on images of sharing: A nation consisted of people with a shared culture, a shared history, and a shared language.
>
> Johnstone (2000, p.409)

But what is it exactly that must be "shared" by members of a particular group for them to feel part of the same nation or, in our case, to be under the impression that they speak the same language: an *ideal*, or a *practice*?

What is interesting about the previous argument of Itkonen (2008, p.279) regarding the English dictionary being about "something that is *common to* or *shared by all* speakers of English" is that it is obviously *false* in practice: no individual English speaker actually holds anything that resembles the content of the full English dictionary in their mental lexicon, and each and every one of those mental lexicons is also "uniquely structured by specific, idiosyncratic linguistic experiences" (Rodd, 2020, p.1) so that all speakers of English probably do not even share a single common mental lexicon to begin with. Under a social and normative ontology of language, what is implied as being "shared" is obviously more a (normative) *ideal* than a *practice* since, as we can see with the argument of Itkonen, it is not the content of speakers' mental lexicons that is shared but the *ideal lexicon* embodied by the public dictionary. As far as normal language is concerned, then, what is shared is not so much what speakers *do* know or how they *do* behave, but rather what they *ought* to know and how they *ought* to behave.

Interestingly enough, such considerations can be made sense of in the context of nationalism, in turn. One could argue indeed that the whole nationalist enterprise should be understood as a project striving to create a sense of community from the sharing of a normative ideal rather than the sharing of a practice *per se*. French people, for instance, would feel French by virtue of sharing a common normative ideal of what it means to be "French"—of how they all *ought* to behave as French people—rather than by virtue of a common practice of being French—of a common way of behaving in the same "French way" indeed. French individuals could then all end up being French *in their own way* without this necessarily compromising the whole French nationalist enterprise, as nationalism would not, in fact, require people to all *behave* exactly in the same way for a sense of community to emerge, but only for them to share a common representation of how they all *ought* to behave. With respect to language, this would mean that speakers could all speak their own idiolect and still be said to speak the "same language" so long as they all shared the same normative ideal of what it meant to speak their language correctly.

Of course, the very idea of a *normative* ideal carries with it the assumption that it should influence the practice at some point: if anything, it remains a "Durkhemian ideal" endowed with some form of authority and so able to *constrain* human behavior eventually. People may not need to behave exactly the same way to feel part of the same nation at the beginning, but sharing a common normative ideal should arguably lead them to progressively align their respective behaviors nonetheless. Yet, it is precisely the underlying sequentiality that matters here—between what always remains *first* a shared ideal and only becomes *then* a shared practice—for it crucially reminds us that nations are always "*imagined* communities" (Anderson, 1983/2016) rather than enactments of pre-existing realities—and as such forms of "sharedness" that we actively *build* rather than merely *label.* As Gellner (1964) puts it indeed:

> Nationalism is not the awakening of nations to self-consciousness: it invents nations where they do not exist [...]
>
> Gellner (1964, p.168)

Such considerations take on their full meaning with respect to language through the question of whether the regularities governing our linguistic behaviors are in fact regularities *de facto* or *de jure.* What is it indeed that we *do* share as far as language is concerned: a common *practice* of language use manifested by regularities *de facto* between our respective linguistic behaviors and so a common specification of how we *do* behave, or rather a common normative *ideal* of language use better characterized as regularities *de jure* between our linguistic behaviors and so a specification of how we *ought* to behave? Understanding the importance of that distinction proves critical to the understanding of how the assumption of shared language and shared meaning actually holds empirically despite the pervasiveness of subjectivity. For when we are presented with empirical evidence in support of the shared meaning assumption, what we are presented with most of the time is actually elicited agreement that pertains to speakers' appreciation of the *norm* and not to their respective linguistic *behaviors*—in other words, agreement on the *ideal* and not alignment of the *practices.* To better illustrate my argument here, let me take a concrete example which will allow me to recap some of the critical arguments I have been carrying out so far in this section—notably with respect to the myth of homogeneity and the marginalization of subjectivity—as well as connect the present discussion to the considerations of the previous chapters.

Back when I was trying to uncover the theoretical underpinnings of the concept of lexical similarity in computational linguistics, I stumbled across a paper by Mosier (1941) which caught my attention insofar as it explicitly claims to bring experimental evidence in support of the assumption of shared meaning.[100] The paper begins as follows:

---

[100] I started my literature review from (Bruni et al., 2014; Hill et al., 2015; Gerz et al., 2016) and (Rubenstein & Goodenough, 1965; Miller & Charles, 1991) all the way to (Osgood, 1952; Osgood et al., 1957) since Lenci (2018) mentions specifically the pioneering contribution of Osgood (1952) to the field of distributional semantics I was originally interested in:

> Vector-based representations of meaning, like those later adopted in [Distributional Seman-

> The meaning of a word is a relationship between an individual, a symbol, and the object, event, or relationship symbolized. This meaning is affected, moreover, by the context of stimuli, symbolic and otherwise, in which the relationship occurs. It is supposed, in this study, that what we customarily refer to as the "meaning of a word" varies, not only from one word to another, but from one individual and from one context to another. As an example of this last, we have "*only fair*," "*fair work*," and "*fair maidens*." In spite of the variation in the meaning of a given word due to the individual and to the context, there appears to remain a residuum of meaning which acts as if it were inherent in the word (not, of course, in the word *per se*, but in the accumulation of social experiences which for any word, are relatively the same from one individual and from one context to another). Thus while the word "*beautiful*" may have different meanings for different persons, and different meanings when applied to different objects, it seldom, if ever, means the same as is meant by "*ugly*." Moreover, the variations which do occur seem lawful and relatively small in magnitude.
>
> Mosier (1941, p.123)

Note, first, how Mosier fundamentally acknowledges the intrinsic subjectivity of word meaning—"the 'meaning of a word' varies [...] from one individual [...] to another"—but only to immediately marginalize it (quantitatively) right after—"the variations which do occur seem lawful and relatively small in magnitude". What is more, the rationalization of this marginalization is explicitly grounded in the myth of homogeneity, since Mosier considers that "the accumulation of social experiences [...] for any word, are relatively the same from one individual [...] to another". The singularity of our socio-linguistic experiences is thus explicitly denied—or at the very least marginalized—since Mosier also emphasizes later on in his paper that:

> The social nature of language is such that words tend to be presented to different individuals (within a homogenous language group) in approximately similar situations.
>
> Mosier (1941, p.124)

Mosier's (1941) experimental protocol requires human subjects to rate individuals words based on their degrees of "favorableness" and "unfavorableness". The written instructions to the participants go as follows:

> Here are a number of words often used to express your judgment about a person, an object, or an event. These judgments range from extremely favorable

---

tics], were pioneered in psychology by Osgood (1952), who defined the semantic system as a semantic space of n-dimensional feature vectors representing concepts (Lenci, 2018, p.152)

I found (Mosier, 1941) to be cited in (Osgood, 1952; Osgood et al., 1957) as well as in (Henley, 1969), itself cited in (Miller & Charles, 1991).

> to extremely unfavorable. [...] While these words are not all applicable to the same object, they do have in common the quality of favorableness or unfavorableness of the judgment expressed by the word. Thus, while "hateful" might not be applied to the same object as would "repulsive," both are alike in expressing unfavorable judgments. It is this *favorableness* or *unfavorableness* of judgment which you are asked to make the basis of your rating.
>
> Mosier (1941, pp.126–127)

But what will interest us here more particularly is the remark that he adds right after the specification of those original instructions, where he tells us that:

> These written instructions were supplemented by verbal instructions, found necessary after a few trial administrations: that the basis of the rating should be the meaning the subject understood as he read the word in a book, rather than the meaning which he might intend as he used the word [...]
>
> Mosier (1941, p.127)

I find it particularly interesting to put this remark into perspective with what Carston (2002) tells us more than 60 years later:

> I see no reason to think we are completely bereft of intuitions about what the words in our language mean, though these intuitions are easily mixed with intuitions about standard usage.
>
> Carston (2002, p.169)

For what is it indeed that Mosier is actually testing with this idea of the meaning of a word "as [...] read [...] in a book", if not a "standard usage" that might as well be referred to here as the "normal meaning" of a word corresponding to the normative ideal of how it *ought* to be used rather than how it *is* actually used by individual speakers in practice—since Mosier explicitly suggests that this *normal meaning* should directly oppose how speakers might individually intend to use the word? If so, then it tells us everything we need to know about this purportedly demonstrated "shared meaning" which, if anything, should be but a shared norm—a shared normative *ideal*, in fact—and not a shared behavior and so not a shared *practice*. What is more, the fact that Mosier felt the need to specify those additional instructions after a few trials suggests that the experimental protocol originally failed to support the hypothesis of shared meaning. In fact, it actually brings support to the intrinsic subjectivity of word meaning, for it means that when left without experimental instructions constraining their judgments so as to specifically target the normative ideal of normal meaning, subjects naturally mobilize their "subjective meanings" which appear to significantly differ from one another.

Be that as it may, Mosier's (1941) results still suggest that *something* is shared as far as meaning is concerned—even if only a normative ideal of normal meaning—which is certainly a significant result. Significant, yes, but *for whom*? Why should this result be relevant to the psychological or language sciences at large? After all, the fact that people *do* agree on a norm does not necessarily tell us that they actually *follow* it in practice—otherwise, why would have Mosier needed to specify those verbal instructions in the first place? So the first question to ask ourselves is whether this normal meaning actually intervenes within the interpretation process in any way, which is far from given.

Mosier's opposition between "the meaning the subject understood as he read the word in a book" and "the meaning which he might *intend* as he used the word" (my emphasis) anticipates the fundamental Gricean distinction between *conventional meaning* and *intentional meaning* (see Section 1.4.1). That distinction, as we have seen, can easily be made sense of within the ideal/practice dichotomy so long as one approaches conventions are regularities *de jure* and not *de facto*. The opposition between *ideal* and *practice* thus evokes the opposition between *objective* and *subjective*, between *public* and *private*, but also—as the conventional/intentional dichotomy indeed suggests—between *sentence meaning* and *speaker meaning*, and between *what is said* and *what is meant*.[101] Normal/conventional meaning is thus (under the Gricean approach at least) considered to play a central role in the interpretation process. It is, in fact, presumed to constitute the systematic point of departure from which all interpretations are derived. However, this assumption has been the specific target of Clark (1997) and is part of what he calls the "dogmas of understanding"—and is targeted more specifically by his "dogma of sentence meaning" (Definition 38) and his "dogma of saying" (Definition 39).

> **Definition 38 (Dogma of sentence meaning)**
> For listeners to understand what a speaker means, they must first determine the meaning of the sentence uttered.
>
> Adapted from Clark (1997, p.569)

> **Definition 39 (Dogma of saying)**
> For listeners to understand what a speaker means, they must first determine what the speaker is saying.
>
> Adapted from Clark (1997, p.572)

It has also been the target of the whole cognitive pragmatics enterprise which argues that interpretation is much more inferential than conventional and emphasizes notably that not every utterance (or word, for that matter) even has a conventional or normal meaning to begin with (see Levinson, 1983, p.17).[102] So the fact that this purportedly shared

---

[101] Remember that Grice (1975) originally characterized his employment of the word "say" as follows:

> In the sense in which I am using the word *say*, I intend what someone has said to be closely related to the conventional meaning of the words (the sentence) he has uttered. (Grice, 1975, p.44)

[102] We may thus wonder whether the experimental results of Mosier are not to be put on the account of an experimental setup comprising a limited and biased sample of tested lexical items—296 he tells us,

normal/conventional meaning could contribute anything to our understanding of human language and language process is actually far from obvious.

If speakers do not process normal meaning and if nobody actually speaks normal language, can we really expect language models to contribute anything to our understanding of human language and cognition? (see Dupre, 2021; Piantadosi, forthcoming, for recent discussions on that point).[103] Well, can you learn anything about the human lexicon by looking into the public dictionary? This may sound like an easy take, but in any case the contribution of this section lies elsewhere, and notably in the contextualization of the ontology of normal language. Part of the reasons why normal language is never quite made explicit indeed is that it also rests on an everyday concept of "language" that, as such, often feels unnecessary to explicate. Normal language is, in a sense, "just" language—as if the concept itself was obvious and self-explanatory—yet, as the considerations of this section critically suggest, normal language is very much socio-historically and socio-culturally situated. Contextualizing normal language as I did in this section should naturally lead us to question the universality of its underlying assumptions. As far as shared language is concerned, what we can say is that it is not obvious that every socio-linguistic community necessarily shares a common normative ideal of linguistic behavior. Johnstone (2000) mentions notably the work of Le Page & Tabouret-Keller (1985) on that point (see also Le Page & Tabouret-Keller, 2010, for a brief overview) and tells us that:

> [I]ndividuals' models of linguistic norms are less consistent in more heteroglot, culturally diverse settings, where people are less likely to project linguistic self-images similar to those of their neighbors.
>
> Johnstone (2000, p.416)

And it is not obvious that sharing a common normative ideal of linguistic behavior is even necessary to enable successful communication in the first place. Johnstone (2000), citing the work of Dorian (1994) this time, tells us indeed that Gaelic speakers of East Sutherland, Scotland, seem to perfectly manage communication without linguistic norms:

> [S]ome variation in Gaelic speech simply does not take on social meaning: One form is just as good as another. This means that individuals are freer to speak differently from one another.
>
> Johnstone (2000, p.416)

Looking back at the results of (Mosier, 1941), we may also wonder if they could not be considered the product of a biased sample of subjects in addition to a biased sample of tested items. Mosier does not tell us much about the participants to his study beyond the fact they were 140 students in introductory and second courses in psychology (see Mosier,

---

including duplicates (see Mosier, 1941, pp.125–126, for details)—"biased" to the extent that they would so happen to possess a consensual normal meaning.

[103] As you will have understood, I would argue that language models speak normal language and that they process normal meaning by extension.

1941, pp.127–128, for details). But we can reasonably assume that they come from his institution of affiliation (the University of Florida) or an English-speaking academic institution at the very least, most likely located in the USA.[104] This biased sample of subjects would certainly question the generalizability of his results (see Henrich et al., 2010a,b, for a general perspective) but they could also very much explain them. Indeed, what if the apparent agreement that he elicits from his subjects on normative ideals of normal meaning was not the product of his experimenting exclusively on subjects evolving in particularly normative socio-linguistic environments, governed by all sorts of linguistic norms notably with respect to what words are supposed to mean? Come to think of it, what is more normative that the academic environment which constantly reminds students that words ought to be used is a “correct” way with their “correct” meanings—and which sanctions them when they deviate from those norms of correctness? Mosier’s results would then appear for what they are: an artifact of an experimental setup biased by its focus on a specific community of human beings structured around a normative ideal of linguistic behavior that its members are in fact very much educated to share. Those are all the considerations I propose we bear in mind when I say that normal language is *shared* language.

3.3.3 Smart machines

In his *Introduction* to *The Turing Test*, Shieber (2004a) comments briefly under a footnote to Turing’s (1950) original question “Can machines think?” that:

> Turing used the terms “think” and “be intelligent” as if they were synonyms, as one can tell by a simple comparison of his article’s title and first sentence. In common usage, the two often mean quite distinct things. When I say that my son is intelligent, I usually mean something beyond the fact that he is capable of thought. However, I and many authors follow Turing’s practice, taking the notion of “being intelligent” under which it means “being capable of thought”, rather than “being smart”.
>
> Shieber (2004a, footnote 2, p.6)

Shieber’s comment above introduces two fundamental comparisons underlying Turing’s conception of intelligence: the *equivalence* between “being intelligent” and “being capable of thought” and the *opposition* between “being intelligent” and “being smart”. In the first comparison, the important keyword is not so much the term “thought” (which suggests that intelligence is conflated with thinking) than the word “capable” which suggests that intelligence is primarily approached as a *faculty*, that is, as a species-level (maybe even species-*specific*) ability to “think” that one possesses by virtue of being human. Such a conception of intelligence is what I propose to call the “humanist conception of intelligence” and which I will refer to in this section as “intelligence” plain and simple so as to remain consistent with Turing’s terminology (Definition 40).

---

[104] As is the case with most US psychology experiments then and now (see Arnett, 2008).

Definition 40 (Intelligence)
Intelligence is an ability to think that one possesses by virtue of being human.

The second comparison then—the opposition between "being intelligent/capable of thought" and "being smart"—points at the fundamental opposition between this humanist conception of intelligence and a normalist conception of intelligence which underlies our everyday use of the term—a conception which I will refer to in this section as "smartness", following Shieber's (2004a) own terminology (Definition 41).[105]

Definition 41 (Smartness)
Smartness is a normative ideal of correct and exceptional human behavior.

The opposition between those two conceptions of intelligence crystallizes notably around the fact that the normalist conception of intelligence presupposes "intelligence" to be something one can have *more* or *less* of—which the humanist conception precisely does *not*. Under the humanist conception of intelligence indeed, intelligence is a universal human ability that one possesses by virtue of being human and which, as such, remains fundamentally incommensurable across members of the same species. You cannot be "more" or "less" intelligent for the same reason that you cannot be "more" or "less" human: you either are, or you are not—you either possess that ability to "think", or you do not.[106]

Such considerations imply two things. First, that Turing (1950) takes "being intelligent" to be essentially synonymous with "being human". Second, that his humanist conception of intelligence remains fundamentally at odds with a normalist conception of intelligence that dominates both our everyday understanding of the term and the field of psychology at large (see Sternberg, 2020, for an overview). This is actually what Fostel (1993) is precisely pointing at in his original critique of the Turing test when he tells us that:

> [The Turing test] is testing humanity, not intelligence.
>
> Fostel (1993, p.8)

---

[105] As we saw in Section 3.2.2.4, normalism proceeds by establishing a normative point of reference that ultimately serves to *rank* and *order* a particular collection of phenomena, items or entities. With smartness, the collection of entities under consideration is the set of *human behaviors* and the normative point of reference in question is *the Galtonian ideal of exceptionality* (by opposition to the Queteletian ideal of normality). Just like normal language, smart behavior does not correspond to the normal/average behavior of a normal/average human being, but to the correct/exceptional behavior of an ideal/exceptional human being (remember Table 3.1 from Section 3.2.1.4 with the example of perfect and imperfect spelling distinguishing smart and intelligent behavior). Every human behavior is thus ultimately ranked according to this normative ideal of smartness—each being more or less "close" to this normative point of reference—so that each human being can ultimately be characterized as more or less "smart".

[106] Understand here that, according to those definitions, typical psychometric tests of intelligence such as IQ tests (Braaten & Norman, 2006) should be understood as tests of "smartness" rather than tests of "intelligence" *per se*. The intuition is plain: to every task in such tests corresponds a set of ideal "correct" responses that characterizes the normative point of reference from which all subject behaviors is ultimately assessed. This normative ideal of correctness characterizes a normative ideal of exceptional human behavior, since no human being should be expected to answer *all* those questions correctly (and even if one does, the test would probably be refactored anyway, since otherwise that would mean that no human being could ever be consider *more* intelligent (i.e. "smarter") than this threshold).

Indeed, "intelligence" in (Turing, 1950) is not what we are commonly used to: it is not a normative ideal of correct and exceptional human behavior that ultimately serves to rank us in relation to one another, but a characterization of what we all purportedly share as human beings and that makes us all equally human.

Now, of course, approaching intelligence as a faculty and a universal endowment of the human species does not mean that it is necessarily readily available to every human being from birth. Like other biological faculties—such as *language* or *vision*—it can be conceived as an "organ" that must grow and develop—notably in interaction with its environment and through exposure to external stimuli—before it can turn into an *effective* ability manifesting itself through observable *behavior* (see Anderson & Lightfoot, 2000, on approaching the language faculty as an organ). Until then, it would remain but a mere "potential"—as Turing (1948/2004) himself makes clear in his earlier writings:

> [T]he potentialities of the human intelligence can only be realised if suitable education is provided.
>
> Turing (1948/2004, pp.431–432)

Between *intelligence* and *intelligent behavior* thus lies the fundamental contribution of *development*, which is why the two cannot really be considered perfectly equivalent, even for Turing (1950).[107]

Be that as it may, the Turing test still focuses exclusively on intelligent *behavior*, which is precisely why I have chosen to conflate the definition of smartness above with its behavioral characterization—i.e. to be "smart", as per Definition 41, is nothing more than to be able to exhibit *smart behavior*. As far as the considerations of this section are concerned indeed, the opposition between *intelligence* and *smartness* reduces to the opposition between *intelligent behavior* and *smart behavior*. But what constitutes "intelligent behavior" then? For Turing (1950), as we have seen, "being intelligent" is synonymous with "being human", so does that mean that "intelligent behavior" is synonymous with "human behavior", in turn? Not exactly, and it is in fact the very contribution I wish to make in this section. My argument indeed is that Turing equates *intelligent behavior* with *normal human behavior* rather than *human behavior* strictly speaking.

There are at least two ways to make sense of this concept of "normal human behavior" here. The first one is through the normal/exceptional dichotomy and so by opposition to a concept of *exceptional human behavior* which characterizes a normative ideal of "correct behavior" that very few people, if any, actually abide to in practice (see Section 3.3.3.1).

---

[107] Note how, in his above quote, Turing (1948/2004) employs the term "education" rather than "development"—a term which evokes a rather socio-culturally and socio-historically situated practice in comparison (it seems much less straightforward to talk about "human education" indeed than to talk about "human development"). Interestingly, Turing's terminological choice suggests implicitly that what we take to be "intelligent behavior" could very much be dependent on the "education" we have received, and so ultimately to the particular group of human beings we belong to (a question that will prove central to Section 3.3.3.2). Incidentally, it also points at what constitutes a crucial question for the developmental sciences at large, in effect whether the usual characterization of "human development" really encompasses the full spectrum of human behavior (see, e.g., Forbes et al., 2022, for a discussion).

The Turing test fundamentally requires us to build machines that specifically deviate from this normative ideal of correctness characteristic of exceptional human behavior. Indeed, it requires us to build machines that "make mistakes"—which is, in fact, precisely the point that his critiques opposed to Turing with respect to his approach to machine intelligence, and which he refers to in his original paper as the "Arguments from Various Disabilities":

> (5) *Arguments from Various Disabilities.* These arguments take the form, "I grant you that you can make machines do all the things you have mentioned but you will never be able to make one to do W". Numerous features W are suggested in this connexion. I offer a selection:
>
> > Be kind, resourceful, beautiful, friendly [. . .] have initiative, have a sense of humour, tell right from wrong, **make mistakes** [. . .] fall in love, enjoy strawberries and cream [. . .] make some one fall in love with it, learn from experience [. . .] use words properly, be the subject of its own thought [. . .] have as much diversity of behaviour as a man, do something really new [. . .]
>
> Turing (1950, p.447, my emphasis in bold)

As we will see in Section 3.3.3.1, the opposition between *intelligent behavior* and *smart behavior* reduces to an opposition between two different normative ideals of human behavior (Table 3.3), which gives rise to two difference scientific projects for the construction of artificial minds: *artificial intelligence* on the one hand, or the modeling of normal human behavior corresponding to the original project of Turing; and *artificial smartness* on the other, or the modeling of exceptional human behavior corresponding to the current project of modern systems such as ChatGPT (remember Table 3.2 from Section 3.2.2.3).

Machine behavior	Human reference
Intelligent behavior	Normal human behavior
Smart behavior	Exceptional human behavior

Table 3.3: Intelligence against smartness, or normality against exceptionality.

The second way to make sense of this concept of "normal human behavior" then is through the interpretation of the normal as the *familiar* (see Section 3.3.3.2). This interpretation of the normal—which I have yet to introduce—refers to that paper by Dupré (1998) on *Normal People* where he tells us that:

> There is, no doubt, an unsophisticated usage according to which what is normal is what is familiar, and the unfamiliar is feared or condemned as abnormal. But since we are all sophisticated this need not detain us.
>
> Dupré (1998, p.221)

Yet, ‘unsophisticated” as it may be, this interpretation of the normal proves particularly useful to us here as it helps us make sense of judges’ intuitions in the Turing test—as we will see throughout Section 3.3.3.2. For indeed, the Turing test is built on an apparent paradox: it rests, in some sense, on the fundamental unreliability of human intuition as to what constitutes human behavior. In order for a machine to successfully pass the test, a human judge must necessarily make a mistake and misidentify the human participant.[108] This has lead some researchers, such as Hayes & Ford (1995), to argue that:

> [The Turing test is fundamentally] a test of the ability of the human species to discriminate its members from mechanical imposters.
>
> Hayes & Ford (1995, p.974)

They argue notably, *contra* Fostel (1993), that:

> [W]hatever that quality is [that the Turing test is testing] it cannot be characteristic of humanity, since many humans would fail a Turing Test. Since one of the players must be judged to be a machine, half the human population would fail the species test.
>
> Hayes & Ford (1995, p.974)

The apparent paradox disappears, however, if you understand that human judges do not ground their discrimination between human and machine participants based on what they take to be *human behavior*, but on what they take to be *normal human behavior*. That is, they do not judge what is *human* but what is *normal*. And since, at the individual level, the normal is the familiar, then we understand how judgments of “humanity” may vary significantly from one judge to the next—they merely reflect the singularity of judges’ respective experiences of the familiar, and so of what is normal *to them*:

> [S]ubjectivity plays a big part in attributing ‘humanness’ to another on the basis of responses to chosen questions.
>
> Warwick & Shah (2015, p.11)

Be that as it may, the Turing test still intends to provide a characterization of *human* intelligence ultimately, so the challenge for it is to move past this first appreciation of the normal as familiar and towards a characterization of the normal able to span across the full spectrum of human behavior. It achieves this by being a *statistical* test, for in practice what is considered “human” is never determined by a single judge but always by

---

[108] The unreliability of human judgment as to what constitutes human behavior is even more salient in the so-called “viva-voce” configuration of the Turing test which involves only a single participant (as opposed to two in the traditional test, see Turing, 1950, p.446, for the original reference). This phenomenon of misidentification of human participants by human judges in the *viva voce* configuration of the Turing test is what Shah & Henry (2005) call the “confederate effect”, and which they analyze more extensively in (Warwick & Shah, 2015).

the aggregate of the judgments of a full *jury*. This statistical nature of the test implies a statistical conception of intelligence, in turn, so that for Turing, *intelligent behavior* is actually *statistically normal human behavior* in the end. This statistical conception of normality then necessarily commits us to a science of the *probable* over the *possible* and of the *normal mind* over the *human mind*, which is why, all things considered, I will argue that Turing remains but a normalist after all.

3.3.3.1 Normal against exceptional, or intelligence against smartness

The tension between (artificial) *intelligence* and (artificial) *smartness* is present throughout the literature on artificial minds, albeit under different forms and using different terminologies. Russell & Norvig (2020), for instance, tell us that:

> Historically, researchers have pursued several different versions of AI. Some have defined intelligence in terms of fidelity to *human* performance, while others prefer an abstract, formal definition of intelligence called **rationality**—loosely speaking, doing the "right thing."
>
> Russell & Norvig (2020, p.31)

In the above quote, we find already some of the critical considerations I have introduced so far, notably the original conception of artificial intelligence of Turing (1950) dedicated to the study of how people *do* behave through this notion of "fidelity to human performance", or the normativity of smart behavior itself through this notion of "doing the right thing" referring to how people *ought* to behave.[109] The tension between artificial intelligence and

---

[109] For Russell and Norvig too, smart behavior is a *normative ideal* of human behavior—an ideal from which "real" human beings should most likely deviate—insofar as they tell us that:

> We are not suggesting that humans are "irrational" in the dictionary sense of "deprived of normal mental clarity." We are merely conceding that human decisions are not always mathematically perfect. (Russell & Norvig, 2020, p.31)

Their reference to "rationality" is interesting to us here, for it reminds us that the scientific study of human behavior is often structured around normative ideals materialized as expectations about how human beings *ought* to behave, normative ideals which are themselves usually grounded in presuppositions regarding human nature. This particular epistemological attitude occasionally leads researchers to underestimate how much people actually deviate from those normative ideals of human behavior in practice, or to systematically approach deviations from those ideals as "errors" caused by "interferences" from other cognitive or biological processes (such as emotions, memory limitations, etc.). As far as *rationality* is concerned, for instance, Kahneman (2011) typically tells us that:

> Social scientists in the 1970s broadly accepted two ideas about human nature. First, people are generally rational, and their thinking is normally sound. Second, emotions such as fear, affection, and hatred explain most of the occasions on which people depart from rationality. Our article challenged both assumptions without discussing them directly. We documented systematic errors in the thinking of normal people, and we traced these errors to the design of the machinery of cognition rather than to the corruption of thought by emotion. (Kahneman, 2011, p.12)

But one could also probably mention the typical example of Chomsky and of his competence/performance distinction in linguistics:

> Linguistic theory is concerned primarily with an ideal speaker–listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such

artificial smartness naturally crystallizes around the question of whether the Turing test constitutes an adequate mode of evaluation for the construction of artificial minds—French (2012), for instance, typically arguing that:

> [W]e need to put aside the attempt to build a machine that can flawlessly imitate humans; for example, do we really need to build computers that make spelling mistakes or occasionally add numbers incorrectly, as in Turing's original article [footnote omitted] in order to fool people into thinking they are human?
>
> French (2012, pp.74–75)

And indeed, "who needs a machine that can't type?" asks the *The Economist* in 1992, in an article titled *Artificial Stupidity* and dedicated to the results of the first edition of the *Loebner Prize Competition* held at the Computer Museum in Boston on November 8 1991 (see Epstein, 1992, for details). The argument—through which Fostel (1993) actually introduces his humanity/intelligence dichotomy—emphasizes the inherent tension between a Turing test that requires machines to "make mistakes" so as to display normal human behavior, and the various applications those machines are usually put to use which precisely require them not to make any:[110]

> The recent Loebner trial [(The Economist, 1992)], in which a computer program achieved considerable success, perhaps passing [the Turing test] as construed for that trial, demonstrates further that [the Turing test] tests for humanity. The computer program used carefully calculated typing errors to fool [the human judge] into believing it to be human. The capacity to make statistically human-like typing errors is not a convincing step towards intelligence yet it seems to be a major step towards passing [the Turing test]. It made the program more human.

> grammatically irrelevant conditions as **memory limitations, distractions, shifts of attention and interest, and errors** (random or characteristic) in applying his knowledge of the language in actual performance. This seems to me to have been the position of the founders of modern general linguistics, and no cogent reason for modifying it has been offered. To study actual linguistic performance, we must consider the interaction of a variety of factors, of which the underlying competence of the speaker-hearer is only one. In this respect, study of language is no different from empirical investigation of other complex phenomena.
>
> We thus make a fundamental distinction between *competence* (the speaker-hearer's knowledge of his language) and *performance* (the actual use of language in concrete situations). (Chomsky, 1965, pp.3–4, my emphasis in bold)

[110] Fostel then anticipates that if artificial intelligence models are unlikely to prove useful for applications requiring artificial smartness, artificial smartness models are unlikely to be able to pass the Turing test:

> Suppose [an alien] robot had scanned and absorbed all of Earth's libraries, media, daily conversations and so on. Would the robot be able to pass [the Turing test]? I think not. [...] the Robot will be so different from any available human [...] that this failure in discrimination by [a human judge] would be unlikely. A nearly omniscient agent [...] would dearly not be like any human [...] Even if the alien robot is extremely capable, and superior to any human in intellectual endeavors, it will fail [the Turing test]. (Fostel, 1993, p.8)

Fostel (1993, p.8)

Many of those considerations have been confirmed empirically, notably through the so-called "practical Turing tests" carried out in 2008 at the University of Reading (Shah & Warwick, 2010a,b), or in 2012 at Bletchley Park (Warwick & Shah, 2014, 2016b).[111] Warwick & Shah (2016a), for instance, typically mention how the presence of spelling mistakes is used by judges to discriminate between human and machine participants:

> It is interesting here that the judge did correctly identify the human entity as there were a lot of spelling mistakes in their discourse and the conversation was quite stilted.
>
> Warwick & Shah (2016a, p.1001)

> [T]he occasional spelling mistake seems to add human credibility.
>
> Warwick & Shah (2016a, p.1003)

What is also interesting about those empirical results is that they illustrate how the Turing test still targets a *normative ideal* of human behavior—even if that normative ideal is a normative ideal of *normal human behavior* and not exceptional or correct human behavior this time. For indeed, the Turing test does not exactly require human participants to behave as they would naturally behave outside of the conditions of the test, but to behave according to what they think (or what they think the judge thinks) a normal human being ought to behave under such conditions. So when Warwick & Shah (2015) formalize instructions to human participants for them not to be misidentified as machines by human judges, it is not surprising to see them tell human participants:

> Do not show that you know a lot of things—the judge may conclude that you are too clever to be human. [...] Do not add new material of a different nature even if you feel this is helpful, otherwise the judge may feel that you know too much to be human.
>
> Warwick & Shah (2015, p.11)

This example illustrates not only that the Turing test is a test of intelligent behavior rather that smart behavior, but also that this "intelligent behavior" constitutes a normative ideal of normal behavior under the conditions of the test—and that the Turing test primarily tests how judges think normal people *ought* to behave rather than how they *do* behave in practice.

Be that as it may, when this distinction between artificial intelligence and artificial smartness is acknowledged by the literature, it is usually framed in terms of "usefuleness" and as a difference between a "theoretical" and a "practical" project—suggesting thereby, even if only implicitly, that one (artificial smartness) should be considered more "useful"

---

[111] See (Warwick & Shah, 2016a), for a comprehensive review.

than the other (artificial intelligence). Russell & Norvig (2020), for instance, explicitly tell us that:

> Few AI researchers pay attention to the Turing test, preferring to concentrate on their systems' performance on **practical** tasks, rather than the ability to imitate humans.
>
> Russell & Norvig (2020, p.1836, my emphasis in bold)

Else Bringsjord & Govindarajulu (2022), in their encyclopedic entry on *Artificial Intelligence*, insist themselves on the fact that:

> [M]ost AI researchers and developers, in point of fact, are simply concerned with building **useful**, **profitable** artifacts, and don't spend much time reflecting upon the kinds of abstract definitions of intelligence explored in this entry (e.g., What Exactly *is* AI?).
>
> Bringsjord & Govindarajulu (2022, p.64, my emphasis in bold)

However, framing the distinction between artificial intelligence and artificial smartness in terms of absolute usefulness is somewhat misleading. First of all, because nothing is ever intrinsically "useful" in and of itself. In effect, there is nothing about artificial smartness that makes it *inherently* more "useful" or "profitable" than artificial intelligence—everything depends on what one needs for the purpose at hand. Fostel (1993) himself, following his claim that the Turing test is a test of "humanity" rather than "intelligence", explicitly tells us that:

> This does not mean the [Turing test] is useless. It might be a useful marketing tool for 21st century firms producing domestic robots. No one is going to want a vacuum cleaner that speaks better English than they do. Talking alarms in cars have already earned considerable ire by noting minor errors their human occupants make in closing doors. Harnad's [Total Turing test, see (Harnad, 1992)] is an even better candidate for that marketing function: who wants a baby-sitting robot that makes mom or dad look bad?
>
> Fostel (1993, p.8)

It it thus very much possible (at least theoretically) that intelligent machines could prove equally if not *more* "useful" than smart machines for some specific use cases.

Second, framing the opposition between artificial intelligence and artificial smartness in terms of (intrinsic) "usefulness" is misleading insofar as it obfuscates the fact that what is "useful" is always both *time* and *place*-dependant—and a matter of subjective appreciation as well. But it also obfuscates more fundamentally that smartness *itself*—and so what constitutes the normative ideal of correct and exceptional human behavior—is both time and place dependant, and so both socio-culturally *and* socio-historically situated. Indeed, as the scientific literature on the standard normative conception of "intelligence" has emphasized on countless occasions (e.g. Vernon, 1965, 1969/2013; Berry, 1972, 1980; Sternberg, 1984, 1985, 2004), the normative ideal of smartness can vary significantly from one socio-cultural context to the next, to the extent that what we take to be "smart people" in one context may not be considered so in another, and conversely:

> Consider, for example, a person who is deficient in the ability to negotiate a large-scale spatial environment. Such people are often referred to as lacking a good "sense of direction." Although they can usually navigate through old, familiar terrain with little or no difficulty, they may find it difficult to navigate through new and unfamiliar terrains. To someone who comes from a sociocultural milieu where people spend their lives in highly familiar environments, such as their hometown plus a few surrounding towns and cities, the idea of large-scale spatial navigation would never enter into the conception of intelligence, and such navigation would be an essentially unknown cognitive skill. Navigation in unfamiliar spatial terrains would simply be irrelevant to such people's lives, just as the ability to shoot accurately with a bow and arrow is irrelevant to our lives. Were such navigation to become relevant in the sociocultural milieu, then what is "intelligent" [*smart*] would change for that culture. In the Puluwat culture, for example, large-scale spatial navigational ability would be one of the most important indices of an individual's adaptive intelligence [(Berry, 1980; Gladwin, 1970/2009; Neisser, 1976)].
>
> Sternberg (1984, p.272)

Even more interesting about Sternberg's argument is that it stresses that the normative ideal of smartness we live by always varies through time and that it can be influenced by the abilities of machines themselves:

> One need not go to exotic cultures to find effective differences or changes in what constitutes intelligent [*smart*] behavior. As Horn [(Horn, 1979)] has pointed out, the advent of the computer seems likely to change what constitutes intelligent performance in our society. For example, numerical calculation was an important part of some intelligence tests, such as Thurstone and Thurstone's [(Thurstone, 1938)] Primary Mental Abilities Test. But with the advent of cheap calculators and ever cheaper computers, the importance of numerical calculation skill in intelligent behavior seems to be declining. Certainly, using numerical calculation as one of five subtests measuring intelligence, or as the sole or main index of number skill, would seem inappropriate today, no matter how appropriate it may have seemed when the Thurstones devised their test or even a few years ago when numerical calculation skill was a central part of people's lives in school and out (balancing checkbooks, keeping track of expenses, and so on). The importance of quantitative expertise to adaptive functioning has probably not changed; but what such expertise consists of may well have changed, at least with respect to the requirements of life in today's

> society. Thus, even in our own culture, we see changes over time, no matter how slow, in what constitutes intelligence. Businesses interested in assessing the intelligence of today's job applicants are much more likely to be concerned about skills in learning to use and in using electronic media, and much less concerned about calculational skills, than they were just a few years ago.
>
> Sternberg (1984, p.272)

Sternberg's considerations above are all the more relevant to us here than Turing (1950) himself typically refers to flawless arithmetics as a prototypical example of smart behavior, notably in his response to the "Arguments from Various Disabilities" where he specifically tells us that machines would have to make *arithmetics* mistakes so as to pass his test:

> The claim that "machines cannot make mistakes" seems a curious one. One is tempted to retort, "Are they any the worse for that?" But let us adopt a more sympathetic attitude, and try to see what is really meant. I think this criticism can be explained in terms of the imitation game. It is claimed that the interrogator could distinguish the machine from the man simply by setting them a number of problems in arithmetic. The machine would be unmasked because of its deadly accuracy. The reply to this is simple. The machine (programmed for playing the game) would not attempt to give the *right* answers to the arithmetic problems. It would deliberately introduce mistakes in a manner calculated to confuse the interrogator.
>
> Turing (1950, p.448)

Such considerations could probably extend to the case of *spelling* as well, since the advent of various spellchecks and machine-based spelling assistants seems to have relegated—or to be progressively relegating, at least—perfect spelling outside of the scope of smart behavior indeed (see Pan et al., 2021, for an overview).[112] The normal/exceptional dichotomy proves particularly useful here, for it allows us to get a better grasp of the phenomenon at stake. It explains notably why a particular behavior (e.g. perfect arithmetics or perfect spelling) ceases to be part of our normative ideal of smartness the moment it becomes accessible to most (notably through the help of machines). The reason is that our conception of smartness is intrinsically tied to an appreciation of exceptionality, so that to be considered "smart" a behavior must remain *exceptional* and so inaccessible to most if not all.[113] It

---

[112] Which does not mean that perfect spelling is no longer valued in itself, of course, but only that it is becoming less and less valued for individuals to demonstrate perfect spelling *by themselves*, since it is becoming possible (or at least easier) to achieve perfect spelling with the help of a spellcheck—just like, before that, it became less and less valued for individuals to demonstrate perfect arithmetic skills by themselves since it became possible for them to calculate quickly and accurately with the help of a calculator.

[113] The example of perfect arithmetics and perfect spelling could probably be extended to the example of encyclopedic knowledge as well, since the advent of internet, Wikipedia and search engines seems to have relegated encyclopedic knowledge outside of the scope of smartness to some extent: can we really call someone "smart" for knowing something that can be googled in 5min indeed?

also explains why, in some sense, what is "smart" is often whatever machines *cannot* (yet) do. In effect, smart machines constantly displace the normative ideal of smartness all the while striving to reach it—in what is but a never-ending process? We can probably expect ChatGPT and other artificial smartness models to displace our normative ideal of smartness once again, but as to whether we will ultimately see an end to that process and be able to say that we have *achieved* artificial smartness the question remains open.

Those points having been clarified, we can now recast the opposition between artificial intelligence and artificial smartness as an opposition between two distinct scientific projects, both dedicated to two different objects of inquiry and best visualized through the comparison between the normal and cumulative normal curve. Artificial intelligence corresponds to the project of modeling normal human behavior (Figure 3.3) while artificial smartness corresponds to that of modeling exceptional human behavior (Figure 3.4). Doing so allows us to understand why artificial smartness models such as ChatGPT are unlikely to ever pass the Turing test—contrary to what recent work suggests (e.g. Biever, 2023; Jones & Bergen, 2024). They target exceptional human behavior, which is precisely

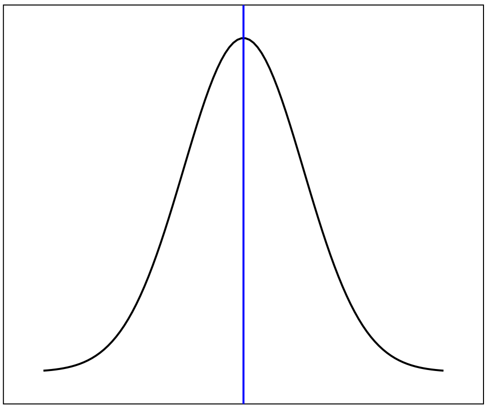


Figure 3.3: Artificial intelligence, or the modeling of *normal* human behavior. Correct machine behavior corresponds to normal/average human behavior marked as a straight blue line at the center of a normal curve modeling a hypothetical distribution over human behavior.

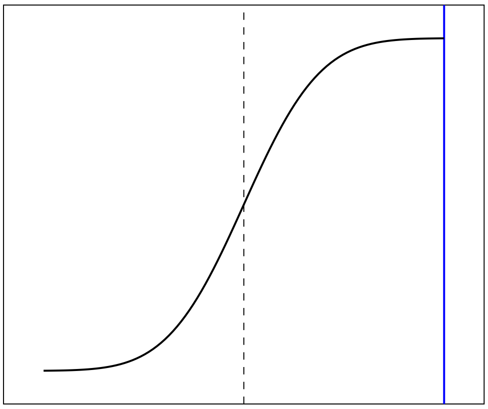


Figure 3.4: Artificial smartness, or the modeling of *exceptional* human behavior. Correct machine behavior corresponds to exceptional human behavior marked as a straight blue line at the far right of a cumulative normal curve modeling a hypothetical distribution over human behavior. Note how the straight blue line marking exceptional human behavior specifically deviates from the normal/average dashed line at the center of the curve.

distinct from the normal human behavior required to pass the Turing test—as the gap between the blue line and the dashed line in Figure 3.4 clearly indicates. This does not mean that such work are necessarily "wrong", however. It is theoretically possible, at least, for an artificial smartness model to pass through a phase where it proves "good enough" to be human in between two phases where it proves "too bad" to be human on one side and "too good" to be human on the other (see Figure 3.5). But as Figure 3.5 suggests, that phase can only be temporary at best, for what is dragging "progress" in the field of artificial smartness is precisely the normative ideal of correct and exceptional human behavior located at the extreme far right of the cumulative normal curve, which necessarily falls outside of the scope of normal human behavior in the end. Be that as it may, the fact that ChatGPT already displays exceptional human behavior such as near-perfect spelling and dictionary-size lexical knowledge suggests that the results of (Biever, 2023; Jones & Bergen, 2024) are most likely circumstantial, and that they should be taken with a grain of salt in any case.

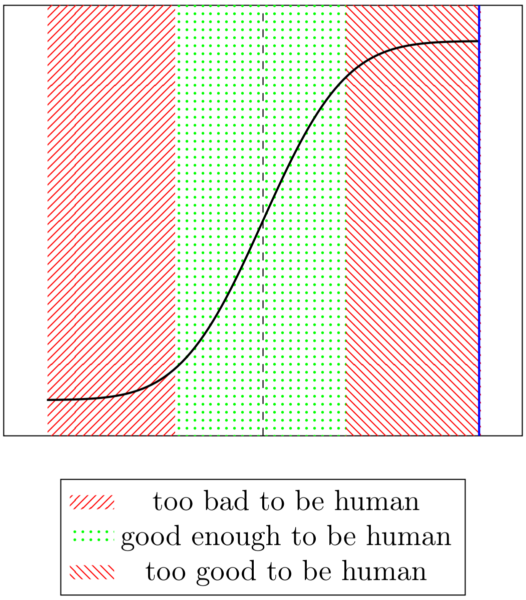


Figure 3.5: Why language models such as ChatGPT are unlikely to pass the Turing test: they are models of artificial smartness where progress is driven by the straight blue line at the extreme far right of the cumulative normal curve (i.e. exceptional human behavior) while passing the Turing test requires targeting the dashed line at the center of the cumulative normal curve (i.e. normal human behavior). Machine behavior is here categorized into three zones: the left red dashed zone where it is too bad to be human, the center green dotted zone where it is good enough to be human, and the right red dashed zone where it is too good to be human. Segmentation between those three zones is completely arbitrary and chosen for illustrative purposes only.

Of course, one could probably argue here that ChatGPT could be tweaked so as to deliberately make (spelling) mistakes and ultimately prove capable of passing the Turing test.[114] However, note that my argument here does not pertain to the capabilities of current language models but to the underlying scientific project they are trying to achieve. My point indeed is to stress that artificial intelligence and artificial smartness are just two scientific different projects which target two different scientific goals and, as such, require two different and irreconcilable criteria of correctness.[115] Naturally, having two distinct criteria of correctness does not preclude that a single machine could be made so as to satisfy them both depending on the context of evaluation—although that would require us first to agree on what we take to be a "single machine" (or the "same model", in the case of language models). In any case, such a machine would require a comprehensive understanding of what characterizes both normal and exceptional human behavior, which in both cases remains far from given—and in some sense constitutes the very purpose of each project.

With all those considerations in mind, we can now better understand why the recent development of "artificial general intelligence" or "human-level artificial intelligence" is not going to get us any closer at passing the Turing test. All those projects constitute mere *extensions* of artificial smartness rather than a return to artificial intelligence in the sense of Turing (1950). Consider how Wang (2019) contextualizes those terminologies:

> The different working definitions of AI correspond to not only different ways to abstract from human intelligence, but also different expectations about the destination of this research.
>
> In the early years of AI research, the works were clearly targeted at computers that are generally comparable with the human mind [(Turing, 1950; McCarthy et al., 1955/2006; Feigenbaum & Feldman, 1963)]. There were ambitious projects like General Problem Solve [(Newell & Simon, 1963)], the Fifth-Generation Computer Systems [(Feigenbaum & McCorduck, 1983)], and the Strategic Computing Program [(Roland & Shiman, 2002)], but none of them reached their declared goal, which led to widespread doubt about the feasibility of the "grand dream of AI," and contributed to the following "AI Winter."
>
> Driven by motivations including to avoid the impossible missions, to obtain the necessary resources, and to improve its public image, the AI community shifted its aim to more realistic tasks, like solving practical problems and carrying out individual cognitive functions. For a long time afterward, topics like "general-purpose intelligence" and "thinking machine" became taboos, and were judged

---

[114] Although note that at the moment, it seems that if you ask ChatGPT to produce spelling mistakes, what you get are not so much "typos" than deviations from a very specific written norm of English, which suggests that ChatGPT treats certain written transcriptions of oral formulations or even certain specific dialectal and sociolectal variations of English as "mistakes" (see Appendix A.3 for examples). Whether or not that would prove sufficient to pass the Turing test is questionable, of course, but what is interesting here is that it also illustrates how "narrow" ChatGPT's appreciation of "correct English" actually is in comparison to the wide diversity of possible English behaviors (see Fleisig et al., 2024, for a relevant discussion on the matter).

[115] The two cannot be reconciled into a single criterion of correctness indeed for if both perfect and imperfect spelling is "correct" as far as machine behavior is concerned, then *anything goes*.

> as not serious or even as pseudoscience. The aim of AI had been degraded to the building of "smart tools" [(Nilsson, 2009)].
>
> In recent years, a renaissance has been happening in AI, partly due to the hope raised by the success of new techniques such as deep learning, and partly due to the realization that the old problems cannot be sidestepped. To distinguish these types of research from the conventional works, new names have been introduced, including "Human-level AI" [(Minsky et al., 2004; Nilsson, 2005; McCarthy, 2007)] and "Artificial General Intelligence" [(Goertzel & Pennachin, 2007; Wang & Goertzel, 2007)]. Some people also call this type of work "Strong AI", though it was not Searle's original meaning of the term [(Searle, 1980)].
>
> Wang (2019, pp.14–15)

In the above quote, Wang explicitly tells us that artificial smartness is historically conceived as a "degraded" form of artificial intelligence—as when he mentions than "the aim of AI had been degraded to the building of "smart tools" ". But as you understand by now, exceptional human behavior is in no way a "degraded" version of normal human behavior. If anything, it is actually the other way around, since normal human behavior is usually framed in terms of "mistakes" deviating from the normative ideal of correct and exceptional human behavior. All we can expect from those models of "artificial general intelligence" is thus to extend the scope of artificial smartness to new domains, but *not* to achieve artificial intelligence in any way. After having achieved perfect arithmetic skills, perfect spelling, and maybe perfect syntax, we may observe new models demonstrating perfect encyclopedic knowledge, perfect logic, or even perfect rationality. But as far as artificial intelligence is concerned, none of those models will be able to pass the Turing test, for they will always model exceptional rather than normal human behavior.

Given all the aforementioned considerations, it is not surprising that artificial smartness models such as ChatGPT do not "think" like humans (remember Section 3.3.1.2). How could such models *function* like humans indeed if they do not even *behave* like humans in the first place? coming to terms with the intelligence/smartness dichotomy also means coming to terms with the extent to which the normative ideal of smartness we live by actually pervades our common appreciation of "intelligence", and how difficult it is to emancipate ourselves from it by when attempting to develop artificial minds. Given the correctness principle indeed, our first reflex always seems to be to align what is "correct" for machines with what is "correct" for humans. Yet, it is precisely what the Turing test is asking us *not* to do.

Back to "true" artificial intelligence now, Turing (1950) specifically tells us that a machine would have to "make mistakes" so as to be able to pass his test. But would *any* "mistake" go? Precisely not, for as he tells us himself in his above quote, the point would be to "deliberately introduce mistakes in a manner calculated to confuse the *interrogator*" (my emphasis).[116] Proudfoot (2013, 2017b, 2020) emphasizes this crucial aspect of the test as well when she tells us that:

---

[116] Distancing ourselves from the specific context of the Turing test for a moment—and notably from the

> *[T]he Turing test does not test machine behaviour.* Instead it tests the observer's reaction to the machine [. . .]
>
> Proudfoot (2017b, p.303)

Passing the Turing test thus depends on *who* evaluates those mistakes, and so on what they take to be normal human behavior. So let me now turn to the normal as familiar.

3.3.3.2 From the normal as familiar to the statistically normal

In his original paper, Turing (1950) defines "success" in his Imitation Game as follows:

> I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent, chance of making the right identification after five minutes of questioning.
>
> Turing (1950, p.442)

On may wonder first why Turing considers a baseline of 70% rather than 0%. The answer to that question is that the Turing test is an *indistinguishibility test* and not a "misidentification test". The point for the machine is not to be systematically misidentified as a human, but only to make it impossible for an average human interrogator to tell it apart from a human. That means that the machine should never be systematically identified as such, of course, but also that it should never be systematically misidentified as a human either. There should simply be nothing "distinctive" in its behavior that sets it apart from a human.

One may wonder then why Turing does not consider a baseline of 50% rather than 70%. A "true" indistinguishability test would indeed suggest that the judgment of an average human interrogator should be based on pure chance—50%, no more, no less. In this case, I would put this specific design choice—alongside that of considering only "five minutes of questioning"—down to the context under which the Turing test was originally conceived. The point was to provide a reasonably challenging goal for a task that was back then considered hard enough already. In effect, a baseline of 70% makes the test slightly "easier" to pass than a baseline of 50%, since with the former the average human interrogator is precisely given "more chance" to make the right guess than with the latter. Those specifications are therefore highly contextual and should not be considered set in stone. In fact, the very spirit of the Turing test would command to update them as artificial

---

question of *normal human behavior*—we also see how, for a humanist approach to intelligence, the fundamental question would be that of whether there exists such things as *impossible* human mistakes—mistakes that no human being will ever make—so that we could truly distinguish *human* from *non-human* mistakes. From a humanist perspective indeed, the investigation into human intelligence is an investigation into all the possible mistakes that we can make, and so into all the possible deviations from the normative ideal of smartness we live by that are made possible by our human nature. The search for human intelligence thus becomes a quest for the boundaries of human subjectivity, understood as our singular deviations from the norm.

intelligence progresses, so we should ultimately consider decreasing the baseline from 70% to 50% and increasing the duration of questioning from 5min to 10min, 30min or even more. Those changes would not alter the nature of the Turing test, just change its level of difficulty.[117]

A no less critical concept in Turing's (1950) specification of success is the notion of "average interrogator" that deserves particular attention. One way to interpret it is by appealing to the normal/exceptional dichotomy and to consider that it opposes the idea of an "exceptional/expert" human interrogator and thus suggests that the test should be reserved to "normal/non-expert" human interrogators only—especially as Turing et al. (1952/2004) explicitly mentioned in a later BBC interview that:

> The idea of the test is that the machine has to try and pretend to be a man, by answering questions put to it, and it will only pass if the pretence is reasonably convincing. A considerable proportion of a jury, who should not be expert about machines, must be taken in by the pretence.
>
> Turing et al. (1952/2004, p.495)

Once again, I would put this specific design choice down to the attempt at building a reasonably challenging Turing test. But note also how, in his original paper, Turing (1950) talks about "*an* average interrogator" (singular) rather than "average interrogator*s*" (plural), suggesting thereby that this notion of "average" applies to the *set* of interrogators as a whole rather than to every human interrogator taken individually. To make sense of this claim here, it is crucial to remember that the Turing test is a *statistical* test—as Turing himself makes clear in his original paper:

> There are already a number of digital computers in working order, and it may be asked, 'Why not try the experiment straight away? It would be easy to satisfy the conditions of the game. A number of interrogators could be used, and statistics compiled to show how often the right identification was given.'
>
> Turing (1950, p.436)

As such, the Turing test never mobilizes a single juror but always a full jury. This notion of "average interrogator" should thus be interpreted as referring to an *abstraction* made of the aggregate of individual interrogators' judgments, rather than to a "real" individual interrogator.[118] The underlying assumption being that the subjective judgments of individual interrogators are always partial and partially erroneous, and that it is only by aggregating those subjectivities together that one can approach the "true" and "objective" value of the target object of inquiry. It is, in fact, the very assumption that we find at the core of most behavioral experiments today, as Cowart (1997) explains:

---

[117] In what follows I will consider an unspecified duration of questioning and a baseline of 50% so as to stay as close as possible from the spirit of a "true" indistinguishibility test.

[118] Note, however, that the two interpretations are not incompatible with one another, and that it is perfectly possible to characterize an "average interrogator" as the aggregate of interrogators all being individually average themselves (in the sense of non-expert).

> In countless psychophysical cognitive and physical tasks, human performance is characterized by a random scatter of individual observations around a more or less stable mean. In statistics, this scatter is termed *error variance* and is ubiquitous in observations of human performance.
>
> [. . .]
>
> In typical psychological studies, error variance is assumed; investigators routinely collect numbers of observations from each participant, often ask each participant to respond to each stimulus many times, and report the procedural and/or statistical methods used to control error variance. Standard practice recognizes that individual observations are not reliable, and sophisticated practitioners may evaluate findings differently according to their confidence in the means employed to control error variance.
>
> Cowart (1997, pp.31–32)

It is an assumption that derives from the very same *metrological realism* that I introduced in Section 3.3.2.3, and which found its way into the behavioral sciences from astronomy notably through Quetelet, who was an astronomer himself.[119] As Rose (2016) explains indeed (see also Desrosières, 1993/1998, especially ch.2):

> Astronomers believed that every individual measurement of a celestial object (such as one scientist's measurement of the speed of Saturn) always contained some amount of error, yet the *total* amount of aggregate error across a group of individual measurements (such as many different scientists' measurements of the speed of Saturn, or many different measurements by a single scientist) could be minimized by using the average measurement [(Simpson, 1755), see also (Hald, 1986)]. In fact, a celebrated proof by the famous mathematician Carl Gauss appeared to demonstrate that an average measurement was as close to a measurement's true value (such as the true speed of Saturn) as you could ever hope to get [(Camic & Xie, 1994; Stahl, 2006) and see also (Sheynin, 1973)]. Quetelet applied the same thinking to his interpretation of human averages: he declared that the individual person was synonymous with error, while the average person represented the true human being [(Quetelet, 1846)].
>
> Rose (2016, p.27)

In sum: the one who determines "success" in the Turing test is not any judge, but the normal/average judge.

With those precisions in mind, we can now look back at previous research and reconsider some of their claims. For instance, when Shah et al. (2012) tell us that:

---

[119] Else see (Tafreshi, 2022) for an explicit account of the influence of Quetelet on modern psychology.

> One feature of 'humanness' that Turing did not factor into his *imitation game* for machine thinking and intelligence is that mistakes will be made by some of the human interrogators, and others are easily fooled.
>
> Shah et al. (2012, p.1)

We can see that this is no exactly true, for "mistakes" in interrogators' judgment is in fact an integral part of the Turing test. First, in the sense that for machines to be truly indistinguishable from human participants, human interrogators precisely *need* to "make mistakes" and misidentify them as humans half of the time on average (with a baseline of 50% at least, less with a baseline of 70%). Second, in the sense that what motivates abstracting away the subjectivities of individual interrogators' is precisely that those subjectivities are presumed to be noisy and so "mistaken" in the first place. It is thus not surprising to observe that, in later work, the very same Warwick & Shah (2015, 2016a) misquote (Turing, 1950) by mentioning "average interrogator*s*" plural rather than "average interrogator" singular, as in:

> Earlier in 1950, Turing had referred to 'average interrogators' rather than a jury and this presents something of a problem in a practical sense if a statistical average of some kind is inferred. The interrogators employed in these tests are certainly interested parties, but other than that they come from a wide variety of backgrounds.
>
> Warwick & Shah (2015, p.2)

Or in:

> To adhere strictly to Turing's wording we have focused here, as best we could, on 'average interrogators' [...] and have included a wide variety of people.
>
> Warwick & Shah (2016a, p.1005)

For everything in their interpretation of (Turing, 1950) actually derives from their considering this notion of "average" to apply only to human interrogators taken individually rather than (also) to the *set* of human interrogators as a whole.

The same considerations hold with Proudfoot's (2017b, p.303) previous quote, then, which suggests (even if only implicitly) that a problem with (or at least a crucial feature of) the Turing test is that it does not rest on an *objective* evaluation criterion but only on the *subjective* appreciation of success by individual interrogators—in that it "does not test machine behaviour" but only "the observer's reaction to the machine". But once again, Turing (1950) does not test the "observer's reaction to the machine" but the *normal/average* observer's reaction to the machine—which makes all the difference. For if it is true that the Turing test *does* start from the subjectivity of individual interrogators, the important point is that *it does not stop there*: it *averages* all the subjectivities together precisely so as to constitute an *objective* criterion for the evaluation of success—as per the standard assumption of metrological realism underlying most if not all behavioral experiments today.

Last, this very same focus on the subjectivity of individual interrogators rather than on the normal/average interrogator is what we found behind the critique of Hayes & Ford (1995) this time, who tell us that:

> The imitation game conditions say nothing about the judge, but the success of the game depends crucially on how clever, knowledgeable, and insightful the judge is.
>
> Hayes & Ford (1995, p.973)

Yet again, as we have seen, success in the Turing test is never determined by a *single* judge but always by the *normal/average* judge which, if certainly unique, remains nonetheless an abstraction constructed from the aggregate of a collective *jury*.

All those references above point at an important blind spot in the Turing test nonetheless. It is true that Turing (1950) tells us nothing about how the jury of human interrogators should be constituted. If the question matters here, it is notably because judges' intuitions as to what constitutes "normal human behavior" appears to be very much socio-culturally situated. Thus, and without any further specification as to what constitutes the population of human interrogators, the guarantee for any Turing test to provide a comprehensive account of *human behavior* at large remains far from granted. A quick detour by the *software engineering stack exchange* forum suffice to illustrate that point. On the thread dedicated to the question "If you could pose a question to a Turing test candidate, what would it be?",[120] user TZHX posts the following response:

> "That September 11th thing was amazing, wasn't it?" – A human would get the reference, a machine is much less likely to.
>
> TZHX, comment posted on April 1st 2011 at 9:27

To which user Job immediately comments:

> I believe that 70% of Earth population [. . .] would not get that reference, and 30% would not get it even if you say it in their native language. We are not being fair to the machines.
>
> Job, comment posted on April 1st 2011 at 15.39

In essence, this example illustrates the argument of French (1990) about nordic seagulls and the Turing test being a test of "culturally-oriented intelligence":

> We will see that the Turing Test could be passed only by things that have experienced the world as we have experienced it, and this leads to the central point of the present paper, namely, that *the Test provides a guarantee not of intelligence but of culturally-oriented* human *intelligence*

---

[120] See `https://softwareengineering.stackexchange.com/q/64248`

French (1990, p.53)

French (1990) parallels the Turing test in a thought experiment aimed at determining whether a machine could really be said to *fly*. He imagines a pool of judges composed exclusively of inhabitants of a large Nordic island for whom, he posits, seagulls would be the only animals known to fly. He concludes that the only machines that would be said to fly under such conditions would be those that basically "fly like seagulls".

Consequently, if we do not pay attention to how the pool of judges is constituted, we may end up equating *intelligent behavior* with the behavior of a *particular* community of human beings—and most likely to simply treat what is "normal" as what is WEIRD and thereby miss the original purpose of the test: accounting for *human* intelligence.[121] However, it is important to understand that the problem at stake is not only a matter of human interrogators, for success in the Turing test is always determined by a triadic relationship between *human interrogators*, *machines* and *human participants*. Train an artificial intelligence model on WEIRD human behavior with WEIRD human interrogators and non-WEIRD human participants, for instance, and chances are that you will make it easier for your machine to be misidentified and so easier to pretend to be normal. But train your model on non-WEIRD human behavior with both WEIRD human interrogators and WEIRD human participants this time, and chances are that you will make it harder for your machine to be misidentified and so harder to pretend to be normal. Of course, none of those situations would characterize "success" in the Turing test anyway. An idealized first case scenario with full machine misidentification would converge to a probability of 0%—i.e. probability for an average human interrogator to make the right guess—while an idealized second case scenario with perfect machine identification would converge to a probability of 100%. Thus, in both cases, we would remain far from the 50% target baseline. But what such considerations fundamentally emphasize is that reports of machine performance on Turing tests cannot be taken at face value and require that we take into account all the parameters involved, and notably what type of behavior the machine was trained on (for a machine learning model at least) and which populations the pools of human interrogators and participants were sampled from. For as we can see, it is always possible to artificially inflate or deflate machine performance on the Turing test *via* certain specific parametric configurations.

So Shah et al. (2012) and Hayes & Ford (1995) above appear to be correct, in some sense: the specifications of the Turing test *do* influence the difficulty of the test. But it is not really

---

[121] Especially given the prevalence of the WEIRD category in traditional behavioral experiments, practical Turing tests included (participants in practical Turing tests are exclusively English speakers, for instance, and mostly people residing in the UK). This is not to say that WEIRD and non-WEIRD people form completely homogeneous categories, of course—let alone that this is the adequate category to employ to best characterize the problem at hand. I am merely using this general category here for its known relevance to the question of sampling bias in the psychological and behavioral literature (see Rad et al., 2018, for a relevant discussion). In what follows, I will consider for the sake of the argument that those categories are irreconcilable with one another and characterize distinct normative ideals of normal human behavior, so that a WEIRD behavior for a non-WEIRD individual is necessarily "abnormal" in comparison to a non-WEIRD behavior and *vice-versa*.

a matter of “cleverness”, “insightfulness”, “knowledgeability” of “foolability” of individual interrogators. In fact, framing the problem as such misses the crucial point, namely, that the Turing test is a *statistical* indistinguishibility test where success is determined not by a single interrogator but the normal/average interrogator who precisely needs to be not smart but *normally smart* on average. Success in the Turing test is thus better framed in terms of *alignment*—and more specifically, as the relative alignment between machine behavior and the average interrogator’s normative ideal of normal human behavior, in comparison to the behavior of the normal human participant. The point for the machine indeed is always to get as close as possible—but no closer—to the normative ideal of the judge in comparison to the behavior of the participant—to appear *as normal* and so both *no less normal* but also *no more normal* that the human participant on average. It is thus not a matter of intellectual abilities of individual judges and there is actually no need to mobilize such notions—which basically boil down to considering that some subjectivities are essentially “better” than others—to explain the phenomena at hand. It is enough to treat all those subjectivities as intrinsically *different*—which is precisely what metrological realism does—without the need to rank them or consider that some are “better” or “worse” than others.

More interesting now about our previous example of WEIRD/non-WEIRD sampled populations is the configuration where both human participants and machines would display WEIRD behaviors, for instance, but with a non-WEIRD human interrogator this time: how would the interrogator evaluate the situation? The best way to get an intuition about the problem at hand—given the specifications I introduced before—is to replace the general WEIRD category by a specific *language*. Imagine, for instance, an English speaking human interrogator confronted with a non-English speaking machine and a non-English speaking human participant—an English speaking interrogator who, for the sake of the thought experiment, would have absolutely no knowledge of any other human language whatsoever. How would that human interrogator determine who is who in this configuration? A possible hypothesis, consistent with statistical thinking, would be to consider that their decision will be based on pure chance. If we generalize this example to the average human interrogator, however, we understand that it implies that the machine would be able to pass the Turing test under this specific configuration, irrespective of the “correctness/incorrectness” of its linguistic behavior. No matter what the machine would say and how it would say it indeed, its behavior will always remain truly indistinguishable from that of the human participant from the perspective of the human interrogator since that human interrogator only speaks English and therefore cannot discriminate between any two non-English linguistic behaviors anyway. This example thus further illustrate that, to be truly *meaningful* as far as the goal of artificial intelligence is concerned, Turing test scores must be reported alongside parametric configurations (and that some form of overlap between human interrogators and participants is probably to be required so as to avoid such edge cases in the first place).[122]

---

[122] By “overlap” here, I mean that human judges and participants should be sampled from the same population and take turn as judges and participants.

Table 3.4 provides a recap of our discussion so far regarding the effects of various parametric configurations on Turing test scores. As we can see, a fundamental challenge for the Turing test is to rely on samples of human interrogators and participants which are truly representative of the human population as a whole. The question of “representativeness” is tricky (see Chasalow & Levy, 2021, for an overview of the concept in the context of statistics and machine learning) and in this case it is all the more so than it actually pertains to the set of possible normative ideals of normal behavior which, if truly grounded in the singularity of our respective experiences, are probably as numerous as human beings on Earth. What is more, grouping some of those normative ideals together proves all the more delicate than the Turing test actually exacerbates those singularities—since in practice both machines and human participants compete to get as close as possible to the normative ideal of the human interrogator—and thus narrows down our appreciation of the normal at the individual level. So let me avoid the question here and consider the very hypothetical case where we would have access to unlimited resources allowing us to include the entire human population in the test, both as interrogators and as participants. What would be the “normal human behavior” that would emerge from this test? The *statistically normal*, of course. That is: the frequent/typical/(statistically) predominant human behavior, as per the sample of human population at hand. In the end, the Turing test thus equates *intelligent behavior* with *statistically normal human behavior*—and so what is “intelligent” with what is *likely* to be human. It thus focuses not so much on what is *possibly* human than on what is *probably* human.

Interrogator	Machine	Participant	Score
W	W	$\overline{\text{W}}$	0%
$\overline{\text{W}}$	$\overline{\text{W}}$	W	0%
W	$\overline{\text{W}}$	W	100%
$\overline{\text{W}}$	W	$\overline{\text{W}}$	100%
W	$\overline{\text{W}}$	$\overline{\text{W}}$	50%
$\overline{\text{W}}$	W	W	50%
W	W	W	?
$\overline{\text{W}}$	$\overline{\text{W}}$	$\overline{\text{W}}$	?

Table 3.4: Recap of Turing test score (defined as the probability of correct machine identification by a normal/average human interrogator) in idealized scenarios, considering the human population to be split into two complementary sets (W and $\overline{\text{W}}$) with two fully distinct and incompatible normative ideals of normal behavior that cannot be evaluated beyond chance by members of the other set. The last two configurations (full row of W and full row of $\overline{\text{W}}$) correspond to situations where the test would provide meaningful test scores but would only evaluate normal W behavior and normal $\overline{\text{W}}$ behavior rather than normal *human* behavior.

Do all such considerations make Turing a *humanist* or a *normalist*? Before answering that question, let me go first through some important clarifications. As we saw in Section 3.2.2.4—and notably through the example of Chomsky (1975, 1986) and his language faculty—adopting a humanist epistemology does not mean having to dispense from any concept of *normality* whatsoever. Within Chomsky's humanist approach to language, for instance, biological normality remains very much present through the question of possible impairments of the language faculty, which can thus be characterized as normal/abnormal in the sense of functional/dysfunctional. Similarly, adopting a normalist epistemology does not mean having to abandon the concept of *human nature* itself. Normalism remains very much compatible with the idea of a human nature universally shared by all members of the human species, especially if this "human nature" is reduced to a mere "potential" manifesting itself behaviorally only through normal development (see Hull, 1986, and notably p.4 and §2 in general, for a critical overview). So the question cannot be settled only by looking at whether each concept is present or absent from the target epistemology. What we need to look at more specifically is which one of the two concepts constitutes the "central organizing concept" of the target epistemology indeed—as Hacking (1990) originally made clear. As far as Turing's epistemology is concerned, it is here that the question of his behaviorism takes on its full meaning. Many authors have suggested indeed that the major contribution of the Turing test is to turn a "philosophical" question into an *operational* one (see Hodges, 1983/2014; Hofstadter, 1985; French, 1990, pp.334;525;53, respectively). This "operationalism" is often equated with a form of "behaviorism" (e.g. Shieber, 2004a, p.8) insofar as it translates concretely as an exclusive focus on machine *behavior* to arbitrate on machine *intelligence*. But for us here, this statement takes on a particular meaning, for it implies that the Turing test shifts its focus from the *faculty of intelligence* to *intelligent behavior* which, as we have seen, corresponds to *statistically normal human behavior*. From there on, Turing's normalism becomes clear: the statistical conception of normality brings in a normative comparativism that makes it possible to rank human behavior with respect to the normative point of reference at hand (the normal/average).[123] What is more, if what is intelligent is what is statistically normal, then a perfectly intelligent machine would be one that is perfectly (statistically) normal and so one that exhibits the most frequent behavioral traits found within the human population. What would this perfectly intelligent machine look like? For one, it would probably speak Mandarin rather than English.

3.3.3.3 Intelligence: a scientific concept?

Outside of his seminal 1950 paper, Turing (1948/2004); Turing et al. (1952/2004) offers us an alternative but no less interesting approach to the concept of "intelligence". In 1948/2004, for instance, he mentions that:

> [T]he idea of 'intelligence' is itself emotional rather than mathematical.

---

[123] The potentially confusing element being that Turing's normalism opposes yet another normalist conception of intelligence—*smartness*—which relies on a different normative point of reference: the correct/exceptional and not the normal/average.

Turing (1948/2004, p.411)

And explains what he means by that in a dedicated section:

> The extent to which we regard something as behaving in an intelligent manner is determined as much by our own state of mind and training as by the properties of the object under consideration. If we are able to explain and predict its behaviour or if there seems to be little underlying plan, we have little temptation to imagine intelligence. With the same object therefore it is possible that one man would consider it as intelligent and another would not; the second man would have found out the rules of its behaviour.
>
> Turing (1948/2004, p.431)

In a later BBC interview, he returns to those considerations by telling us that:

> As soon as one can see the cause and effect working themselves out in the brain, one regards it as not being thinking, but a sort of unimaginative donkey-work. From this point of view one might be tempted to define thinking as consisting of 'those mental processes that we don't understand'. If this is right then to make a thinking machine is to make one which does interesting things without our really understanding quite how it is done.
>
> Turing et al. (1952/2004, p.500)

There are actually two ways to interpret Turing's considerations here. The first way is to return to his original 1950 paper where he explains that the epistemic opacity of the child machine—or the fact that it may not always be possible to understand precisely what it does and how it solves the task at hand—should be considered an architectural *feature* rather than a "bug" in the system:

> An important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside, although he may still be able to some extent to predict his pupil's behaviour. [...] This is in clear contrast with normal procedure when using a machine to do computations: one's object is then to have a clear mental picture of the state of the machine at each moment in the computation. This object can only be achieved with a struggle. [...] Most of the programmes which we can put into the machine will result in its doing something that we cannot make sense of at all, or which we regard as completely random behaviour.
>
> Turing (1950, pp.458–459)

For what Turing's (1948/2004; 1952/2004) considerations suggest indeed is that this epistemic opacity—or the "blackbox" nature of artificial intelligence systems today (see Nefdt, 2020, §3, and references therein for an overview)—might as well prove to be what makes those systems *intelligent* in the first place.

But another way to interpret Turing's considerations is to return to the *Chomsky–Foucault debate* (Chomsky & Foucault, 2006) and to Foucault's distinction between "scientific concepts" and "peripheral concepts" in particular, which he introduces as follows:

> [Foucault:] Let's take the example of biology. You will find concepts with a classifying function, concepts with a differentiating function, and concepts with an analytical function: some of them enable us to characterize objects, for example that of "tissue"; others to isolate elements, like that of "hereditary feature"; others to fix relations, such as that of "reflex." There are at the same time elements which play a role in the discourse and in the internal rules of the reasoning practice. But there also exist "peripheral" notions, those by which scientific practice designates itself, differentiates itself in relation to other practices, delimits its domain of objects, and designates what it considers to be the totality of its future tasks. The notion of life played this role to some extent in biology during a certain period. [...] I would say that the notion of life is not a *scientific concept*; it has been an *epistemological indicator* of which the classifying, delimiting, and other functions had an effect on scientific discussions, and not on what they were talking about.
>
> Chomsky & Foucault (2006, pp.5–6)

For what Turing's (1948/2004; 1952/2004) considerations would suggest this time is that intelligence is not so much a *scientific* concept than a *peripheral* concept that primarily serves to designate the object of study and set the horizon of the corresponding scientific project. If so, then, "intelligence" would be, in the end, whatever is left for us to understand as far as human behavior is concerned.

3.4 Towards a subjectivist science of human being

In the last two sections of this thesis, I will take the opportunity to put all the considerations I have developed so far into perspective within a more general philosophy of language and biology. In Section 3.4.1, I will return to some critical questions regarding the validity of the Turing test for constructing a science of human intelligence—questions which I have had to leave out so far for reasons of space—and address them through a comparison between *intelligence* and *bipedalism*. In Section 3.4.2, I will then return to the question of language more specifically, and discuss how my subjectivism should be interpreted within a broader philosophy of linguistics.

3.4.1 Towards a science of structure without order

3.4.1.1 Machine bipedalism

Suppose that, inspired by Turing's (1950) paper on *machine intelligence* which asks "Can machines think?", I were to write myself a paper on *machine bipedalism* asking "Can machines walk?"—and by that you will have understood that I mean *walk "like humans"*. Suppose that, inspired by Turing once again, I were to argue that this question was "too meaningless to deserve discussion" and that, following his example, I were to devise my own "Imitation Game"—titled for the occasion the "footprint test"—so as to operationalize the question at hand and determine whether a machine could reasonably be said to "walk" (like humans). The footprint test would consist in having two participants—a human and a machine/robot—leave two separate trails of (bare) footprints on a sandy beach, and a human judge having to determine, based solely on the observation of those trails, which one was left by the human and which one was left by the machine. If, on a repeated number of occasions, the trail of the machine could not be distinguished from that of the human, then the machine could be said to have successfully passed the test.

Critiques of my footprint test would surely not miss the opportunity to criticize its "behaviorism" by arguing notably—following what Searle (1980) did for the Turing test—that passing the test alone should not constitute a *sufficient* proof of (human) bipedalism since bipedalism is clearly not reducible to the ability to demonstrate *bipedal behavior*. And in some sense they would be right: encyclopedic entries on *bipedalism* such as (Napier, 2019), for instance, typically characterize human bipedalism as a *dual* property consisting in both a special pattern of *behavior*—i.e. the bipedal "striding gait" of (Napier, 1967)—and a specific *morphology*, consisting itself in:

> [...] the rebalancing of the head upon the neck, the characteristic S-shaped curvature of the back, the broadening of the pelvis and the straightening of the legs [...]
>
> Ingold (2004, p.316)

That morphology is famously illustrated by Huxley's (1863/2009) picture in Figure 3.6—a picture which Ingold (2004) is also careful to contextualize for us as follows:

> [T]he picture has been deliberately constructed to tell a story, one that has entered the textbooks and been retold on countless occasions ever since. It is the story of how man's eventual achievement of upright posture laid the foundations for his pre-eminence in the animal kingdom, and for the growth of culture and civilization. In the picture, man marches confidently into the future, head high, body erect, while the stooping apes trundle along obediently behind [...]
>
> Ingold (2004, pp.316–317)

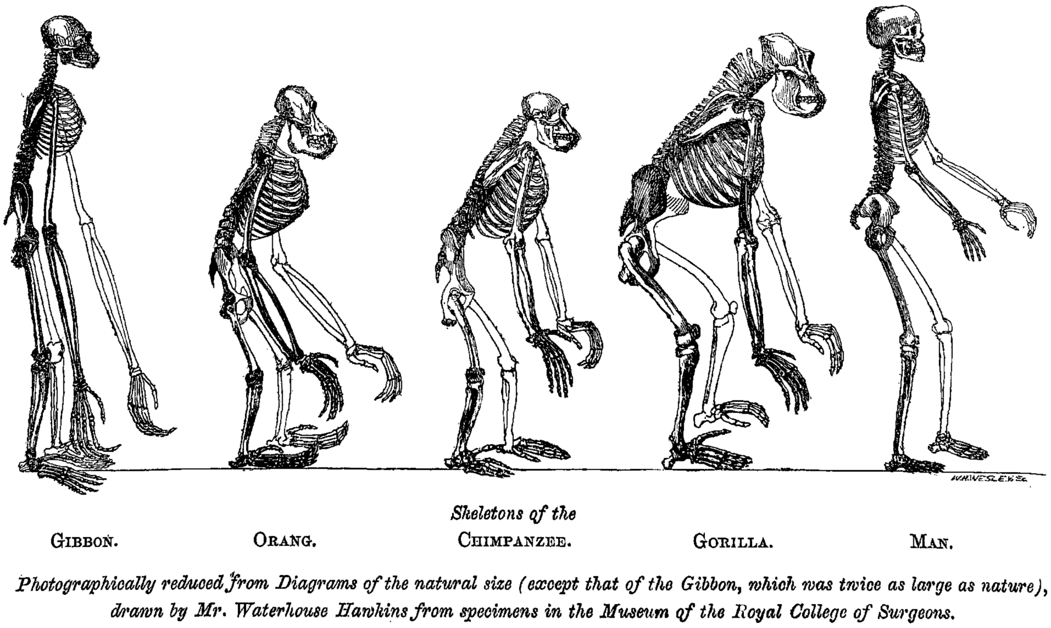

W.H.WESLEY Sc.
Skeletons of the
GIBBON.
ORANG.
CHIMPANZEE.
GORILLA.
MAN.
Photographically reduced from Diagrams of the natural size (except that of the Gibbon, which was twice as large as nature), drawn by Mr. Waterhouse Hawkins from specimens in the Museum of the Royal College of Surgeons.

Figure 3.6: Comparative illustration of skeletons of primates, from (Huxley, 1863/2009).

Could I then, following what Turing (1950) did on machines "thinking", argue that within a few years time, "the use of words and general educated opinion will have altered so much that one will be able to speak of machines [*walking*] without expecting to be contradicted"? Ultimately, as Chomsky (1996) points out:

> These are questions of decision, not fact; decision as to whether to adopt a certain metaphoric extension of common usage.
>
> Chomsky (1996, p.19)

It is not even clear what such terminological disputes would bring to the questions at hand:

> There is no answer to the question whether airplanes *really* fly (though perhaps not space shuttles). Fooling people into mistaking a submarine for a whale doesn't show that submarines really swim; nor does it fail to establish the fact. There is no fact, no meaningful question to be answered, as all agree, in this case. The same is true of computer programs, as Turing took pains to make clear in the 1950 paper that is regularly invoked in these discussions. Here he pointed out that the question whether machines think "may be too meaningless to deserve discussion," being a question of decision, not fact, though he speculated that in 50 years, usage may have "altered so much that one will be able to speak of machines thinking without expecting to be contradicted"—as in the case of airplanes flying (in English, at least), but not submarines swimming.

> Such alteration of usage amounts to the replacement of one lexical item by another one with somewhat different properties. There is no empirical question as to whether this is the right or wrong decision.
>
> Chomsky (1996, p.19)

In any case, it seems that we *do* talk about "bipedal robots" or "bipedal walking robots" already (e.g. Chevallereau et al., 2009; Mikolajczyk et al., 2022) so there may be little matter left for discussion as far as bipedalism is concerned.

Other critiques of my footprint test would probably tackle its "culturally-oriented" nature this time and argue—following what French (1990) did for the Turing test—that passing the test should not even constitute a *necessary* condition to guarantee (human) bipedalism, for failing to exhibit the type of bipedal behavior we are most familiar with should clearly not mean that one is failing to demonstrate *human* bipedal behavior altogether. Once again, they would be right in some sense, for anthropologists have for long shown that the human bipedal striding gait is affected by socio-cultural factors so that:

> The striding gait would be better viewed as a biologically based cultural trait that predominates as a locomotor form in our own society but one that should not be applied as such to other groups in an ethnocentric way.
>
> Devine (1985, p.551)

As Devine (1985) elaborates indeed:

> Everyone is aware of subtle differences in gait that exist between individuals. One of the first experiences of early visitors to non-Western societies was an immediate realization that the manner of walking of the people they were encountering was quite dissimilar from their own.
>
> Devine (1985, p.553)

Note, however, that one does not have to leave the Western world to observe socio-cultural and socio-historical differences in gait. The anthropologist Marcel Mauss (1934/2007), for instance, in his *Techniques of the Body*, reflects upon his experience as a French soldier in Word War I where he observed how cultural differences among British and French infantries prevented them for marching together in a coordinated fashion. He also reflects upon his experience in New York City, where he observed how certain walking styles would later propagate to Paris, France through Hollywood movies (or so he assumes).[124]

How would I, then, placed under the position of a judge, evaluate a trail of footprints coming from a human bipedal behavior I am not familiar with, such as one characteristic of the bipedal behavior historically associated with native populations of North America throughout the eighteenth and twentieth century, and consisting mostly in a pattern of:

---

[124] More recently, Ebersbach et al. (2000) notice a systematic difference of gait velocity between subjects residing in Tyrol and subjects residing in Berlin, and Lieberman et al. (2010) detail how (running) gait can be influenced by whether or not one wears shoes.

> [. . .] stepping on the ball of the foot (or the whole surface of the foot) rather than rolling from heel to toe; [. . .] pointing the toes straight ahead or slightly inward, rather than outward; and [. . .] placing the feet along a single straight line under the body's center of gravity, rather than placing them at hip width and shifting weight from side to side.
>
> Ranalli (2019, pp.89–90)

Would I, with my antagonistic socio-cultural experience of gait, treat it as the most likely human trail—let alone a human trail *tout court*? That would probably depend on the machine trail itself, of course, but it would also illustrate how—just like with the Turing test—judgments on the footprint test are fundamentally grounded in the normal as familiar and bounded by the limits of one's own subjectivity.

3.4.1.2 Beyond normality and internal order

> Normality is a very slippery notion. It also has had a long history of abuse. Responsible authorities in the past have argued in all sincerity that other races are degenerate forms of the Caucasian race, that women are just incompletely formed men, and that homosexuals are merely deviant forms of heterosexuals. The normal state for human beings is to be white, male heterosexuals. All others do not participate fully in human nature. That white, male heterosexuals make-up only a small minority of the human race did not give these authorities pause.
>
> Hull (1986, p.7)

The question of the representativeness of the normal (and lack thereof) certainly constitutes a major challenge for the epistemology of normalism and if anything, the considerations from Hull (1986) above illustrate how difficult it can be for one to conceive the normal beyond the familiar. By contrast, they also illustrate how the statistical conception of normality can appear so attractive in this context: it constitutes a promise of *objectivity*, which specifically proposes to bypass the limits of human subjectivity by displacing the normative ideal of normality from the (subjectively) *familiar* to the (objectively) *typical*.

Be that as it may, determining what constitutes the proper characterization of the normative ideal of normality is but one aspect of the problem. Figuring out where to operate the segmentation between the *normal* and the *abnormal* is yet another, and not the least: from what point should we consider indeed that a particular deviation from the normative ideal of normality falls under the scope of the "abnormal"? The question has been the object of intense debate, especially given the historical "narrowness" of the concept of normality itself—as the quote from Hull (1986, p.7) above critically suggests. Those debates have lead to the progressive expansion of the scope of normality, recasting what was once considered *abnormal* to what is now simply *different*, and potentially *equally normal*. Today, such debates crystalize notably around the question of *autism* and the "neurodiversity" movement—as its name clearly indicates—which targets more specifically a biostatistical concept of normality intrinsically tied to normative considerations and often too prompt to disqualify the *atypical* as necessarily *dysfunctional* if not *pathological* (see Singer, 1999; Dyck & Russell, 2020; Botha & Cage, 2022; Zamzow, 2023, for details). Such a dynamic of expansion is necessarily bound to stop at some point, however, for in any case it can never extend to the human as a whole by definition. What would be the point of mobilizing of concept of "normality" that would just equate the "human" indeed? Normality works in pair, as its very purpose is to draw a segmentation between two categories (the normal and the abnormal) *within* the collection of entities under study. Without the possibility to characterize certain humans or human traits as "abnormal", normality would thus loose its *raison d'être*. From the point of view of normalism, *all* human beings cannot be normal.

What is more, the specification of a normative point of reference is just one step in a process which ultimate purpose is to generate *internal order* within a collection of phenomena, items or elements under study. As I said in Section 3.2.2.4, normalism proceeds through a normative comparativism which consists in ranking the elements of the collection under study with respect to the normative point of reference at hand—which may or may not be the normal, in fact, as we saw with the typical example of smartness. But why should we need such an internal order as far as the study of human intelligence is concerned? Why should we need a normalist conception of intelligence in the first place? Why should we need to rank human behaviors with respect to one another to understand what distinguishes human from non-human intelligence, for instance? This necessity is far from obvious—despite what French (1990) seems to suggest in his critique of the humanist conception of intelligence:

> If we agree that the underlying neural mechanisms are essentially the same across species, then we ought to treat intelligence as a continuum and not just as something that only humans have. It seems reasonable to ask a good test for intelligence to reflect, if only approximately, those differences in degree. It is especially important in the study of artificial intelligence that researchers not treat intelligence as an all-or-nothing phenomenon.
>
> French (1990, p.56)

Consider indeed the counter example of bipedalism: it seems perfectly possible to treat bipedalism as a "continuum" across species—at least insofar as we can talk about "optional" or "facultative" bipedalism in other primates (see Hutchinson & Gatesy, 2001)—without having to consider that any human being is necessarily "more" or "less" bipedal than another. What is more, bipedalism is clearly a non-exclusively human biological trait, since we count birds, penguins, certain marsupials and rodents (such as kangaroos and jerboas) and many mesozoic reptiles as bipeds too (see Hewes, 1961, p.688, for details).[125]

---

[125] We also count bears as "optional bipeds" outside of non-human primates, for instance.

The thing with "intelligence" however—in comparison to *bipedalism*—is that it is more often than not mobilized to explain a difference *across humans* rather than *across species*.[126] The fact of the matter is that this internal order has historically been used to justify corrective measures against those that were considered to deviate too far from the normative ideal at hand (e.g. Carlson, 2017; McConnell & Phelan, 2022), or to naturalize particular *social* or *racial* orders in turn (see Cave, 2020; Adams, 2021; Blili-Hamelin & Hancox-Li, 2023, and references therein for a comprehensive overview within the context of artificial intelligence). Take IQ tests, for instance. Blili-Hamelin & Hancox-Li (2023) specifically tell us that:

> The first IQ test, developed by Binet and Simon in 1905, was intended to help institutions identify students with learning difficulties, for the purpose of separating them from students of "normal" intelligence.
>
> Blili-Hamelin & Hancox-Li (2023, p.274)

And Adams (2021) further stresses that:

> Such tests went on to be used in the US and British colonies to justify the idea that intelligence was a hereditary attribute largely endowed to the white race [(Sehlapelo & Terre Blanche, 1996; Tilley, 2011; Laher & Cockcroft, 2014)].
>
> Adams (2021, p.188)

But we might as well return to Galton himself here, since his opposition between normality and exceptionality is precisely what underlies the standard normalist conception of intelligence (or smartness). As far as the normalist conception of intelligence is concerned, ethical and scientific considerations are always fundamentally intertwined but in the case of Galton, MacKenzie (1981) explains that the former can clearly be said to *precede* the latter insofar as:

> [T]he needs of eugenics in large part determined the content of Galton's statistical theory.
>
> MacKenzie (1981, p.52)

As he further details:

> Galton's work clearly legitimated the elevated position of the professional élite to which he belonged. His argument was that it was a natural élite, not merely a social one. A chapter of *Hereditary Genius* [(Galton, 1869, pp.37–49)] was devoted to justifying the twin propositions that the most eminent were exceptionally well-endowed by nature, and that there were no overwhelming barriers

---

[126] Hence the fundamental difference between the *normalist* approach to intelligence and the *humanist* approach to bipedalism.

> to the most able achieving eminence, even if they were born outside the élite. So membership of the professional élite was deserved, and those outside it had no reason to complain. People like him—he included both his own and his wife's families amongst his examples of the inheritance of ability—were innately superior.
>
> MacKenzie (1981, p.53)

So we see what is at stake here: the point for Galton is ultimately to legitimize a particular social order by arguing that it rests on a biological and so in some sense "natural" order of things and beings. The epistemology of normalism is thus here mobilized to rationalize an internal order among human beings and in a way that makes sense, for if your ultimate purpose is to rationalize a particular hierarchy among human beings indeed, your first step will probably be to find a way to rank them in relation to one another.

Creating an internal order among human beings thus appears not only unnecessary as far as the study of human intelligence is concerned but also very much suspicious if we put it into historical perspective. In any case, it should probably be treated with caution given the dreadful consequences that it has had—and keeps having—on our communities. Ultimately, however, it should be possible to conceive a science of human being without having to rely on normative ideals of human behavior or on normative comparisons between human beings, and we should be able to account for the fact that two human beings can be *different* without having to argue that one is necessarily "better" than the other.

3.4.1.3 Beyond human nature and external order

> Generations of philosophers have argued that all human beings are essentially the same, that is, they share the same nature, and that this essential similarity is extremely important. Periodically philosophers have proposed to base the essential sameness of human beings on biology. In this paper I argue that if 'biology' is taken to refer to the technical pronouncements of professional biologists, in particular evolutionary biologists, it is simply not true that all organisms that belong to *Homo sapiens* as a biological species are essentially the same. If 'characters' is taken to refer to evolutionary homologies, then periodically a biological species might be characterized by one or more characters which are both universally distributed among and limited to the organisms belonging to that species, but such states of affairs are temporary, contingent, and relatively rare. In most cases, any character universally distributed among the organisms belonging to a particular species is also possessed by organisms belonging to other species, and conversely any character that happens to be limited to the organisms belonging to a particular species is unlikely to be possessed by all of them.
>
> Hull (1986, p.3)

There is certainly something attractive from a scientific perspective in approaching bipedalism as a *faculty*, that is, as a universally shared endowment of the human species—even if only a shared *potential*. For one, it allows us to separate the question of human nature from that of individual behavior, and to account for the fact that a faculty may be shared by a species as a whole without it necessarily being behaviorally manifest in each and every one of its individual members. Baby humans, for instance, can very much remain "bipeds" despite their not being able to demonstrate bipedal behavior (yet). They are just bipedal beings *to be*, sharing with other human beings a faculty/potential for bipedalism which should become behaviorally manifest ultimately provided normal/functional development.

But what about those who may never develop that ability to demonstrate bipedal behavior or who may loose it at some point for whatever reason? Can they still be said to retain that "potential" for bipedalism in any meaningful sense of the term and if not, can they still be considered "bipeds" nonetheless? This question invariably triggers in me a feeling of discomfort—a feeling which appears to be shared with others since user Chris H, for instance, replies to the question "Is a human amputee still a biped?" on the *English language and usage stack exchange* by saying:[127]

> I certainly wouldn't want to make the argument that physical disability (or even drunkenness) makes someone less human.
>
> Chris H, (partial) answer posted on March 27th 2019 at 17:44

It almost feels like questioning one's bipedalism is tantamount to questioning their humanity—and for a good reason: bipedalism has always been considered a critical part of our human nature; a fundamental aspect of what makes us all human. Hewes (1961), for instance, begins his paper on the origins of human bipedalism by stating that:

> Man's upright posture and bipedal gait have been justly considered outstanding features of human nature [. . .]
>
> Hewes (1961, p.687)

Ingold & Vergunst (2008) begin their own ethnographic review of walking by asking:

> When did our walk begin? When will it ever end? We cannot remember, and will never know. Walking, in this regard, is much like talking, and both are quintessential features of what we take to be a human form of life.
>
> Ingold & Vergunst (2008, p.1)

And didn't Plato himself famously define human being as a "featherless biped" in the first place?

Such considerations rest on the fundamental assumption that human nature is necessarily *shared*, so that every trait or character that compose it must be universally distributed

---

[127] See `https://english.stackexchange.com/questions/491596/is-a-human-amputee-still-a-biped`.

across all members of the human species. Yet, as Hull (1986) crucially emphasizes, this assumption remains fundamentally at odds with the basis of Darwinian biology, for indeed:

> If evolutionary theory has anything to teach us it is that variability is at the core of our being.
>
> Hull (1986, p.6)

As he elaborates further:

> From the evolutionary perspective, all alleles which we now possess were once more than just rare: they were unique. Evolution is the process by which rare alleles become common, possibly universal, and universally distributed alleles become totally eliminated. If a particular allele must be universally distributed among the organisms belonging to a particular species (or at least widespread) in order to be part of its 'nature,' then natures are very temporary, variable things.
>
> Hull (1986, p.9)

Far from being anecdotal, such considerations illustrate the fundamental contribution of the Darwinian revolution to our very conception of "species":

> Since Aristotle, species have been paradigmatic examples of natural kinds with essences. An essentialist approach to species makes sense in a pre Darwinian context. God created species and an eternal essence for each species. After God's initial creation, each species is a static, not an evolving group of organisms. Darwinism offers a different view of species. Species are the result of speciation. No qualitative feature—morphological, genetic, or behavioral—is considered essential for membership in a species. Despite this change in biological thinking, many philosophers still believe that species are natural kinds with essences.
>
> Ereshefsky (2022, p.3)

Hull (1986) traces back the necessity to characterize human nature as a shared essence to ethical and moral considerations:

> [W]hy is it so important for the human species to have a nature? One likely answer is to provide a foundation for ethics and morals.
>
> Hull (1986, p.11)

Which allows him to point at the core of the problem:

> All the ingenuity which has been exercised trying to show that all human beings are essentially the same might be better used trying to explain why we must all be essentially the same in order to have such things as human rights. Why must we all be essentially the same in order to have rights? Why cannot people who are essentially different nevertheless have the same rights? Until this question is answered, I remain suspicious of continued claims about the existence and importance of human nature.
>
> Hull (1986, pp.11–12)

In the above paragraph, Hull (1986) gives us clues to understand the nature of my discomfort upon questioning one's bipedalism. It derives from the intuition that denying one's bipedalism would not only make that person "less human", but that it would also question more fundamentally that person's claim to have rights. As such, it tells us something crucial about the essentialist approach to human nature: it still rests on a certain conception of "order", but an *external order* this time which applies not *within humans* but *across species* and puts human being on top of the hierarchy of species—alone among those entitled to have rights.[128] Essentialism is thus a *speciesism* in some sense, a form of "human supremacism" that frames the singularity of the human species in terms of superiority rather than mere difference from other species.

Be that as it may, it is still true that, by getting rid of internal order, essentialism gets rid of the ethical problems posed by normalism. For if all humans beings are essentially the same indeed, then trying to specify a normative comparison between them becomes somewhat pointless. At the same time, this resolution comes at a cost: a near impossibility to accommodate the fundamental contribution of Darwinian biology, which commands that we place individual variability at the core of our epistemology. This is precisely where subjectivism comes into play. Its focus on individual subjectivity must be put into perspective with Darwin's own methodological individualism—remember Darwin's (1859/2008, pp.37;42) quote from Section 2.2.4 indeed—and its conception of the singularity of every human mind put into perspective with the singularity of every (human) being in general.[129] To be able to safely speak about individual variability and subjectivity, however, subjectivism must first get rid of normative comparisons between those individualities, if it wishes to provide an ethical science of human being, less able to be twisted by supremacist agenda of any kind. To do so, and as we understand from the previous considerations, it must thus get rid of order altogether—which explains why, in the end, it intends to provide a science of *structure without order*.

---

[128] Remember how Ingold (2004, pp.316–317) specifically mentions that Figure 3.6 serves to reinforce the "pre-eminence [of human being] in the animal kingdom".

[129] Which does not mean having to abandon any notion of "sharedness" among human beings whatsoever—just like with the 'Phylogenetic Species Concept", for instance, does not abandon the notion of sharedness among its characterization of species but merely recasts it as a *shared ancestry* rather than a *shared essence* (see Ereshefsky, 2022, pp.19–28, for details).

3.4.2 Towards a subjectivist linguistics

3.4.2.1 Bringing back subjectivity

In the Chomsky–Foucault debate (Chomsky & Foucault, 2006) Foucault explains how Chomsky's humanism in linguistics—which is very much an *essentialism* itself given how it posits the faculty of language to form a common essence shared by all human beings indiscriminately—can be understood as a project consisting in bringing back the *speaker* in a field otherwise dedicated to the study of *language*—a language mainly approached as a social, public and supra-individual entity:[130]

> [Foucault:] [I]n the historical studies that I have been able to make, or have tried to make, I have without any doubt given very little room to what you might call the creativity of individuals, to their capacity for creation, to their aptitude for inventing by themselves, for originating concepts, theories, or scientific truths by themselves. But I believe that my problem is different to that of Mr. Chomsky. Mr. Chomsky has been fighting against linguistic behaviorism, which attributed almost nothing to the creativity of the speaking subject; the speaking subject was a kind of surface on which information came together little by little, which he afterwards combined. [...] Therefore I have, in appearance at least, a completely different attitude to Mr. Chomsky apropos creativity, because for me it is a matter of effacing the dilemma of the knowing subject, **while for him it is a matter of allowing the dilemma of the speaking subject to reappear**. But if he has made it reappear, if he has described it, it is because he can do so. **The linguists have for a long time now analyzed language as a system with a collective value**. The understanding as a collective totality of rules allowing such and such a knowledge to be produced in a certain period, has hardly been studied until now. [...] Here my aim rejoins, with imperfect methods and in a quite inferior mode, Mr. Chomsky's project: accounting for the fact that with a few rules or definite elements, unknown totalities, never even produced, can be brought to light by

---

[130] Gastaldi & Pellissier (2021) carry out the same argument in the context of distributional semantics, where they explicitly connect the distributional hypothesis to the structuralist paradigm:

> As such, a distributional approach is at odds with the generative perspective that dominated linguistic research during the second half of the 20th century. Indeed, the latter intends to account for linguistic phenomena by modelling linguistic competence of cognitive agents, the source of which is thought to reside in an innate grammatical structure. In such a framework, the analysis of distributional properties in linguistic corpora can only play a marginal role, if any, for the study of language. [footnote omitted] By referring the properties of linguistic units to intralinguistic relations, as manifested by the record of collective linguistic performance in a corpus, the distributional hypothesis imparts a radically different direction to linguistic research, where the knowledge produced is not so much about cognitive agents than about the organization of language. It follows that, understood as a hypothesis, *distributionalism constitutes a statement about the nature of language itself*, rather than about the capacities of linguistic agents. (Gastaldi & Pellissier, 2021, p.570)

Note that Scholz et al. (2024), in their typology of linguistic theorizing, incidentally call Chomsky's approach "essentialism" as well (see Scholz et al., 2024, §1.3)

> individuals. To resolve this problem, **Mr. Chomsky has to reintroduce the dilemma of the subject in the field of grammatical analysis**. To resolve an analogous problem in the field of history with which I am involved, one has to do the opposite, in a way: to introduce the point of view of understanding, of its rules, of its systems, of its transformations of totalities in the game of individual knowledge.
>
> Chomsky & Foucault (2006, pp.15–19, my emphasis in bold)

However, Chomsky did not bring back *any* kind of subject: he brought back the *cartesian subject*—with its purportedly universal subjectivity and its claim to find in the retreat to one's own mind access to both knowledge and truth (Avramides, 2020, p.3)—a cartesian subject which had actually become the main target of Wittgenstein's (1953/2009) private language argument before that:[131]

> PLA [The private language argument] directs itself against the dominant tradition of Western philosophy, a tradition equally represented by Descartes, Hume, and Kant. According to this ('Cartesian') tradition, public things and qualities are *reducible to* subjective experiences, which constitute the 'rock bottom' of knowledge. Moreover, knowledge of *other minds* is supposed to be gained on the basis of the 'argument from analogy': When I perceive that bodies (constructed out of my sense-impressions and) resembling mine behave under similar circumstances in the same way as my body does, I may infer with a high degree of probability that these bodies are possessed by minds which think and feel in ways similar to mine.
>
> Itkonen (2008, p.280)

It is thus not surprising to see *introspection* and introspective judgment in general have become such a central methodological instrument within the Chomskyan epistemology. Neither it is to see it have become the main target of a normalist approach to linguistics shifting its focus of analysis from the cartesian speaker to the *normal speaker* if not *normal language* entirely (e.g. Edelman & Christiansen, 2003; Wasow & Arnold, 2005; Featherston, 2007; Gibson & Fedorenko, 2010; Gibson et al., 2013). After all, Durkheim (1901/2013) himself did say that public "things" could not be discovered through introspection alone:

> What indeed is a thing? The thing stands in opposition to the idea, just as what is known from the outside stands in opposition to what is known from the inside. A thing is any object of knowledge which is not naturally penetrable by the understanding. It is all that which we cannot conceptualize adequately as an idea by the simple process of intellectual analysis. It is all that which the mind cannot understand without going outside itself, proceeding progressively by way of observation and experimentation from those features which are the

---

[131] Chomsky makes his affinities with Descartes clear in his *Cartesian Linguistics* (Chomsky, 1966/2009).

> most external and the most immediately accessible to those which are the least visible and the most profound. To treat facts of a certain order as things is therefore not to place them in this or that category of reality; it is to observe towards them a certain attitude of mind. It is to embark upon the study of them by adopting the principle that one is entirely ignorant of what they are, that their characteristic properties, like the unknown causes upon which they depend, **cannot be discovered by even the most careful form of introspection**.
>
> Durkheim (1901/2013, p.7, my emphasis in bold)

3.4.2.2 Subjectivity is not noise

Take Featherston (2007), for instance, who argues plainly that:

> [I]t is no longer tenable for syntactic theories to be constructed on the evidence of a single person's judgements, and [. . .] real progress can only be made when syntacticians begin to think more carefully about the empirical basis of their work and apply the minimum standards we propose.
>
> Featherston (2007, p.269)

Behind his argument, we find the very influence of *metrological realism* I introduced in Section 3.3.2.3, notably through the assumption that individual measurements/judgments are intrinsically "noisy"; that variation across individual judgments should thus be smoothed out by averaging over a large data pool (of informants); and that only this "golden mean" counts as genuine, "true" and objective data:

> Judgements are fundamentally noisy, and show some variability both between informants and across judging events by the same person. It is likely that the second variability explains the first to some extent, that is, that a difference in judgements between two informants is at least partly the result of each individual's judgements being subject to random error. But whether or not this is the case, both these sorts of variability can be evened out, if we obtain multiple independent judgements. The errors cancel each other out and the judgements cluster around a mean, which we can take to be the 'underlying' value, free of the noise factor. Multiple informants thus deliver more accurate data.
>
> Featherston (2007, p.284)

Yet, as the response from Den Dikken et al. (2007) clearly shows, the opposition between humanism and normalism in linguistics with respect to the validity of introspective judgment is not just a methodological dispute. It is an opposition between two approaches to the language sciences—two epistemologies, in fact—which both rely on two different sets of presuppositions:

> In every field of scientific inquiry, it is important to have a clear definition of the object of study and the methodology used to investigate it. For generative grammarians, from the inception of the framework, the object of study has primarily been the speaker's knowledge of his or her native language.
>
> Den Dikken et al. (2007, p.335)

In this case, the opposition translates notably as an opposition between two different ontologies of language—*private* in one case, *public* in the other—which is precisely why humanism *can* rely on introspective judgment to uncover the true nature of language: because for it indeed, language is *not* a (public) "thing". But it also translates as a more fundamental opposition between two antagonist objects of study—the *possible* in one case, and the *probable* in the other:

> Universal Grammar is the abstract grammatical system that, by hypothesis, is innate in the mind/brain of all humans. Every speaker's grammar is, once again by hypothesis, a reflection of Universal Grammar, and, as a result, every speaker's grammar must meet the requirements imposed by the invariant principles and the particular parameter-settings allowed by UG. Given this, relying on the judgments of individual speakers certainly should not be incompatible with studying UG. [...] UG lives in every individual, and there is in principle nothing wrong with trying to model the grammar of a single individual as a way to study its limits and possibilities.
>
> Den Dikken et al. (2007, p.341)

Those antagonist objects of study explain precisely why humanism and normalism adopts two fundamentally opposing approaches with respect to individual variability and why, for humanism in particular, individual variability is *not* noise:

> 'Outliers' should not be cast aside as 'noise' as a matter of course. Whenever an apparent 'outlier' presents itself, one should try to ascertain whether it might be correlated with some other 'outlier' from the same informant. What matters, after all, is what patterns there are in the data. One apparent 'outlier' may team up or correlate with another apparent 'outlier' elsewhere, and the correlation between these 'outliers' may be enormously revealing from a theoretical point of view. But just as one cannot know what 'a group' is before one has encountered it (in the form of judgments that pattern together), one also cannot know in advance what constitutes 'noise'.
>
> Den Dikken et al. (2007, p.350)

As far as humanism is concerned indeed, individual variability is not noise insofar as every linguistic behavior or idiolect in general—singular as it may be—constitutes a *possible* manifestation of the language faculty and, as such, provides crucial empirical evidence to determine its nature and ultimate boundaries. Hence its historical focus on *idiolects* and on the methodological individualism that it entails.

3.4.2.3 From subjectivity to subjectivism

The humanist ontology of language and its corresponding focus on idiolects has been the object of much confusion from the normalist side, however. As when Featherston (2007), for instance, typically argues that:

> The 'my idiolect' definition of the object of study thus buys invulnerability to questions about its data base at an enormous price. It excludes the possibility that there is a universal grammar. It limits the applicability of any analysis to an individual. It is also demonstrably wrong, since studies gathering judgements from groups show unambiguously that judgements of well-formedness are shared across speakers with only limited systematic variation.
>
> Featherston (2007, p.279)

And that:

> The big questions in generative linguistics are those which refer to all speakers, not just one speaker, and to the whole language, if possible to all languages, not just to a single lexical string.
>
> Featherston (2007, p.309)

In this particular case, the confusion crystallizes notably around the question of what constitutes the very nature of that "Universal Grammar", and as Den Dikken et al. (2007) importantly clarify:

> Universal Grammar in Chomsky's I-language sense is not the set of Greenbergian (E-language) universals, that is, the set of surface properties that are common to sets of languages.
>
> Den Dikken et al. (2007, p.341)

In other words, Universal Grammar is not a characterization of all the features that idiolects may share at the surface level, but a characterization of all the constraints that they must satisfy at a deeper level so as to qualify as *human* languages. Universal Grammar is what determines the set of *possible* human idiolects and ultimately constrain their diversity.

Be that as it may, such confusions still capture a fundamental tension behind the humanist enterprise. Indeed, the fact of the matter is that, despite its methodological individualism, it still does not abandon the idea of *shared language* and the idealization of "language" altogether—the only means of "proceeding rationally", Chomsky (1979/2007) even tells us:

> Certainly, it is true that no individual speaks a well-defined language. The notion of language itself is on a very high level of abstraction. In fact, each individual employs a number of linguistic systems in speaking. How can one

> describe such an amalgam? Linguists have generally, and quite properly, proceeded in terms of an idealization: Let us assume, they say, the notion of a homogeneous linguistic community. Even if they don't admit it, that is what they do. **It is the sole means of proceeding rationally**, so it seems to me. You study ideal systems, then afterwards you can ask yourself in what manner these ideal systems are represented and interact in real individuals.
>
> Chomsky (1979/2007, p.54, my emphasis in bold)

Therefore, brining back the cartesian subject within the humanist epistemology actually came at a price: the necessary *quantitative marginalization of subjectivity*, so that the introspective judgment of a single speaker could be extrapolated to all speakers in order to derive a generative grammar of the target (idealized) "language" at hand. Concretely, speakers of the same linguistic community are all presumed to speak the *same language*, share the *same linguistic knowledge* and produce *identical grammaticality judgments*:

> Knowing the language L is a property of a person H [. . .] for H to know the language L is for H's mind/brain to be in a certain state; more narrowly, for the language faculty, one module of this system, to be in a certain state S_L.
>
> [. . .]
>
> The transition from the initial state to the steady state takes place in a determinate fashion, with no conscious attention or choice. **The transition is essentially uniform for individuals in a given speech community despite diverse experience**. The state attained is highly articulated and very rich, providing a specific interpretation for a vast array of sentences lacking close models in our experience.
>
> [. . .]
>
> [T]he basic problem is that our [linguistic] knowledge is richly articulated and **shared with others from the same speech community**, whereas the data available are much too impoverished to determine it by any general procedure of induction, generalization, analogy, association, or whatever.
>
> Chomsky (1986, pp.22;51;55, my emphasis in bold)

So if Chomsky *did* bring back the speaking subject to the field of linguistics, the fact of the matter is that he did not bring back *subjectivity* along with it. If anything, he brought back *inter*subjectivity, in fact—and a *grammatical* intersubjectivity, to be more specific (see Taylor, 1992, pp.66–70, for details)—focusing on what speakers all have in common rather than on what makes them individually singular. Which explains why, in the end, his humanism is *not* a subjectivism but a "universalism", which treats the subjectivity of a single speaker as applicable to all.[132]

However, the quantitative marginalization of subjectivity within the universalist paradigm could only last so long, for it soon became clear that grammaticality judgments could not always be considered identical from one speaker to the next, even among speaker of the "same linguistic community":[133]

> What was not realized for many years was the staggering extent of interspeaker variation on any given set of sentences. It was apparently believed that if one took a set of sentences and elicited judgments about them from some group of speakers, these speakers would agree among themselves as to the degrees of grammaticality of the test sentences. Occasionally, one found asides to the effect that "one dialect," or more accurately, "some speakers," liked (or disliked) one test sentence more than the rest of the subjects did, but for a long time, syntacticians were content to proceed on the assumption that this was a rare

---

[132] As such, work that undertake to confirm the validity of introspective judgment through normalist methods applied to populations of speakers (e.g. Sprouse & Almeida, 2012, 2013, 2017; Sprouse et al., 2013, 2016, 2018) should be taken for what they are: attempts at validating the myth of homogeneity empirically; thereby legitimizing the quantitative marginalization of subjectivity; and ultimately reinforcing the foundational assumption of the universality of the cartesian subject.

[133] This point would have to be nuanced, of course, as I do not mean to imply here that all the experimental results attesting of a certain systematicity across speakers' grammaticality judgments are necessarily wrong. But the question remains: how robust and representative are those results? To what extent are they not a mere artifact of a biased focus on certain syntactic/linguistic phenomena, or a biased sample of subjects? (remember here (Mosier, 1941) from Section 3.3.2.6, and all the literature on representativeness and the generalization crisis in the psychological sciences (e.g. Arnett, 2008; Henrich et al., 2010a,b; Rad et al., 2018; Forbes et al., 2022)). And more generally, to what extent are they not the product of a particular epistemological attitude that precisely requires us to focus on systematicity and similarity (of grammaticality judgments) across speakers, rather than on differences and singularities among them? (remember now (Darwin, 1859/2008, pp.37;42) from Section 2.2.4). Subjectivism is primarily a critique of this particular epistemological attitude, and of the fact that if you posit certain things to be the same, then you are likely to uncover many similarities or identities among them. But change your presupposition now, and posit that they are *not* the same, and chances are you will uncover many differences and singularities among them. Your epistemology is ultimately the pair of glasses you put on to look at the world: it conditions what you see, and what you find. Subjectivism is thus basically asking you to change glasses.

Be that as it may, the quote from Ross below raises an important question for linguistic humanism nonetheless, namely, what should we do with individual differences in grammaticality judgments? To what extent do they question the robustness of this notion of "homogeneous linguistic community" and the idealization of language as a whole? Which amount of individual variation among speakers of "English" is enough to call into question the reasonableness of the approximation it purportedly embodies? And behind that, to what extent can the project of a *generative grammar*—or of *any* idealized "language" grammar for that matter—prove scientifically productive?

Interestingly enough, the various quotes from (Den Dikken et al., 2007, pp.342;350) above and below illustrate that the project of *universal grammar* itself is perfectly compatible with a methodological individualism that would embrace the singularity of idiolects at its core—although it does require a certain reconception of what it means to "generalize" in linguistics. For indeed, the set of latent constraints that universal grammar purportedly characterizes is derived not only from the grammaticality judgments that speakers may form on isolated sentences but also from the *patterns* of judgments they may form depending on the *configurations* permitted by their respective idiolects—e.g., from the fact that speaker A may consider sentence X to be grammatical but not sentences Y and Z while for speaker B it will be the opposite, or yet another configuration. Thus, universal grammar generalizes not only from what speakers have in common at the surface level but also (and more fundamentally) from what singularizes them.

> phenomenon, and could be disregarded, as a first approximation.
>
> Ross (1979, p.128)

More interesting about Ross' (1979) argument is the fact that it makes clear what underlies the quantitative marginalization of subjectivity and the idealization of language:

> After all, it may have been felt, since all these subjects are unquestionably English (Turkish, Mohawk, etc.) speakers, and do not report that they are speakers of different geographical dialects, they have to accept and reject the same sets of sentences, don't they? Otherwise, what could it mean to say "Those people speak the same language"?
>
> Ross (1979, p.128)

As it illustrates, the assumption of shared language constitutes some sort of ontological necessity. The fact of the matter is that we *need* to be able to talk about "shared language" on way or another, and this necessity precedes and conditions the idealization of language and the marginalization of subjectivity. And if we need to do so, it is because of the *presupposition of mutual understanding* once again. Den Dikken et al. (2007) actually make that point clear themselves when they mention how the similarity between idiolects unfolds naturally from the necessity to ensure "perfect mutual intelligibility" between speakers:

> Generalizations over a group can only be made if the individuals that constitute the group share the same linguistic knowledge. Yet there is no a priori reason to expect each individual belonging to a certain group of speakers to have set every single parameter in exactly the same way as his/her fellow group members. Since we know that all human languages change over time, we are forced to conclude that different members of the same 'group' of language speakers internalize subtly different grammars. **Such grammars are obviously very close, ensuring perfect mutual intelligibility**, but may differ in some fundamental ways in their I-language constitutions, thus leading different speakers to analyze a particular string in different ways, with each particular hypothesis leading to different analyses of other input strings and, importantly, to different output utterances regulated by the parameters in question.
>
> Den Dikken et al. (2007, p.342, my emphasis in bold)

This may seem paradoxical for a paradigm that, overall, focuses more on the characterization of shared grammatical structure rather than the shared understanding of linguistic expression (see Taylor, 1992, p.68, for details) and yet even there, as we can see, the presupposition of mutual understanding remains—pervading every aspect of the theory and conditioning some of its critical assumptions, in turn. As if we could not escape it.

Hence comes subjectivism, with its intention to put an end to the historical marginalization of subjectivity—in linguistics and beyond—and why, to do so, it must therefore get rid of the presupposition of mutual understanding altogether. Subjectivism provides an alternative to the standard humanist approach to linguistics that is neither essentialist, nor universalist, but that remains humanist nonetheless insofar as it commits to a science of human nature rather than normal people—even if it conceives this human nature more as a shared ancestry than a shared essence in the end, in line with Darwinian biology and its phylogenetic conception of species.

Not being a universalism means that subjectivism abandons the prospect of a generative grammar, which disappears in any case alongside the myth of homogeneity and the necessity to idealize a shared language, all deriving from the presupposition of mutual understanding. In subjectivism in the end, there are no "languages" beyond idiolects.

Not being an essentialism, it may have to abandon the prospect of a universal grammar as well, though that is not so obvious and does not mean having to throw away the entire contribution of essentialism in any case. Embracing the singularity of linguistic experience and linguistic knowledge remains very much compatible with the methodological individualism that the investigation of Universal Grammar fundamentally requires, though it may lead to an extension or a refactoring of the notion of "generalization" in linguistics once again. Not being an essentialism, however, subjectivism does not posit that this generalization necessarily converges to a single set of constraints, necessarily shared across all human beings. It may as well converge to a scattered set of structures, all singular to each individual, and which only share but a family resemblance with one another across different human beings.[134]

A whole new set of new experiences, of new phenomena, and of new theorizing thus opens up to us at once the moment we accept to abandon the presupposition of mutual understanding. I say it's time we let it go.

---

[134] It may seem paradoxical to argue here that subjectivism is not an essentialism, given that in Chapter 2 I argued that to have a human mind is to be equipped with a common set of primitives and a single shared compositional operator (see p.55). Note however, that for Subjective Coordination Theory those decisions merely correspond to convenient modeling choices and play no particular role in the theory otherwise. More specifically, they do not constitute specific targets of inquiry for the theory and merely corresponds to methodological decisions which purpose is to simplify the computational model overall. That is, the purpose of Subjective Coordination Theory is *not* to determine the nature of those "primitives" or of the compositional operator. Compare that with a linguistic essentialism which purpose is precisely to determine the nature of a language faculty and of a universal grammar purportedly shared by all human beings by virtue of them being human.

Conclusion

This thesis started with an opposition—*language against communication*—and concludes that this opposition is ultimately *a matter of subjectivity.* It is a matter of subjectivity first in the sense that theorizing about language and communication always requires theorizing about subjectivity one way or another as it requires, in particular, arbitrating on the role subjectivity is presumed to play in the overall interpretation process and whether it should be considered *central* or *marginal* to the characterization of language and communication as a whole. But it is also a matter of subjectivity next in the sense that theorizing about language and communication always requires arbitrating among various sets of axioms and presuppositions and depends as such on *your* subjectivity as a researcher and on your personal affinities with particular systems of thoughts. Among those presuppositions, as we have seen, is the crucial *presupposition of mutual understanding* that underlies most if not all of our theories of language and communication today and which conditions the very epistemological attitude we adopt towards subjectivity. I personally stand on the side of those who would like to let it go.

The main contribution of this thesis is to put subjectivity back at the center of our theories of language and communication. It stands in the long tradition of research which attempts to open up new problems by questioning the fundamental assumptions and presuppositions of the old ones. As such, this thesis does not try so much to settle on the question of the communicative function of language, for instance, than to interrogate the terms of the debate and question the very nature of the problem—asking notably *which* communication it is that we are talking about in the first place, and whether that is the *intersubjective/objective* or the *subjective* coordination of minds.

Subjectivity—conceived in its most general sense as the singularity of every being—is thus what fundamentally binds the three chapters of this thesis together. But if I have chosen to employ the term "subjectivity" rather than "singularity" or "individual variability" throughout this thesis, it has been for a specific reason: to make more salient its critical opposition with the concept of *objectivity* that is latently present all along. The necessity of mutual understanding can indeed be said to derive from the necessity of objectivity in some sense, a necessity which is itself threefold:[135]

1. *epistemological*, for how could we possibly do science together if our reflections were not grounded in a shared reality?
2. *political*, for how could we possibly act together if we all lived in incommensurable worlds? and, above all;
3. *metaphysical*, for how could we not all live in the same sensible and perceivable world in the first place?

The first epistemological necessity is the one I attempted to address through my "flat Earth" example at the end of Chapter 1. My point there was to show that subjectivity need not be seen as a *problem* for the practice of science, and that coordinating our subjectivities towards the determination of a shared reality does not necessarily require that we *order* them and consider that some are always more correct, valid, or accurate than others. My point was also to show that we need not oppose our subjective and sometimes contradictory experiences of the world and that we can often reconcile them all within a single theory by asking simply what it is exactly that we have the intuition *of*. Such considerations proved crucial in Chapter 2 to accommodate our intuitions of communication success within Subjective Coordination Theory, which showed that they could very well be accommodated for if recast as intuitions of interaction smoothness rather than intuitions of mutual understanding. They also proved crucial in Chapter 3 to move beyond the apparent unreliability of human judgment as to what constitutes human behavior in Turing tests, which was made possible by recasting judges' intuitions of human behavior as intuitions of normal/familiar human behavior.

The second political necessity is the one I attempted to address in Chapter 2 through the reference to Graeber (2009) and his ethnography of the Direct Action Movement. My point there was to show that subjectivity need not be seen as an obstacle to *action*, as we actually have at our disposable many concrete examples of communities of human beings which structure their democratic practices around subjectivity and the incommensurability of experiences and perspectives and which do not posit that we must necessarily constitute a homogeneous whole (in ideal or in practice) to form community and act together. So even if we never truly understand each other in the end, that should not prevent us from doing things together in any case.[136]

---

[135] I am indebted to Adèle Payet for this formalization of the problem.

[136] Note, once again, that my point in this thesis is not to settle on *whether* we understand one another but on *why* we need to presume that we do so in order to account for our intuitions about language and communication. So what I am implying here is merely that my approach would be compatible with a radical take on language and communication that would posit that we *never* do understand one another in the end—and in fact, I have often been tempted myself to introduce my thesis as such to make matters clearer or more provocative (see below). However, ultimately, its claim is slightly more nuanced. It merely claims that whether we do understand each other or not in practice should not matter, and that we should not have to presume that we do in order to theorize about language and communication.

The last metaphysical necessity, however, is a more delicate matter. It would be an understatement to say that my suggesting we abandon the presupposition of mutual understanding has not always been met with great enthusiasm. A couple of years ago, for instance, I was trying to explain my thesis to my friend Jul and came up with the following: "Basically, I think I am trying to show that people do not really understand each another when they talk". Their reply was immediate, almost instinctive: "How awful!!!". To this instinct of mutual understanding, I can only oppose but my personal experience of calling into question its presupposition. Putting myself in an attitude where I never presume that others necessarily understand me, that we may never share similar views, concepts, meaning or even any form of common ground whatsoever has, I must say, dramatically improved the quality of my relationships. My human relationships have never been so deep, so empowering and so less conflictual indeed than ever since I started presuming that agreement between myself and others could only be but *ad hoc*, superficial and temporary at best, and remained in any case but an *appearance* of agreement which needed to be constantly renegotiated through the confrontation of our subjectivities and the explicitation of their idiosyncracies—a process which, in itself, requires the fundamental acceptance of our intrinsic singularities to begin with.

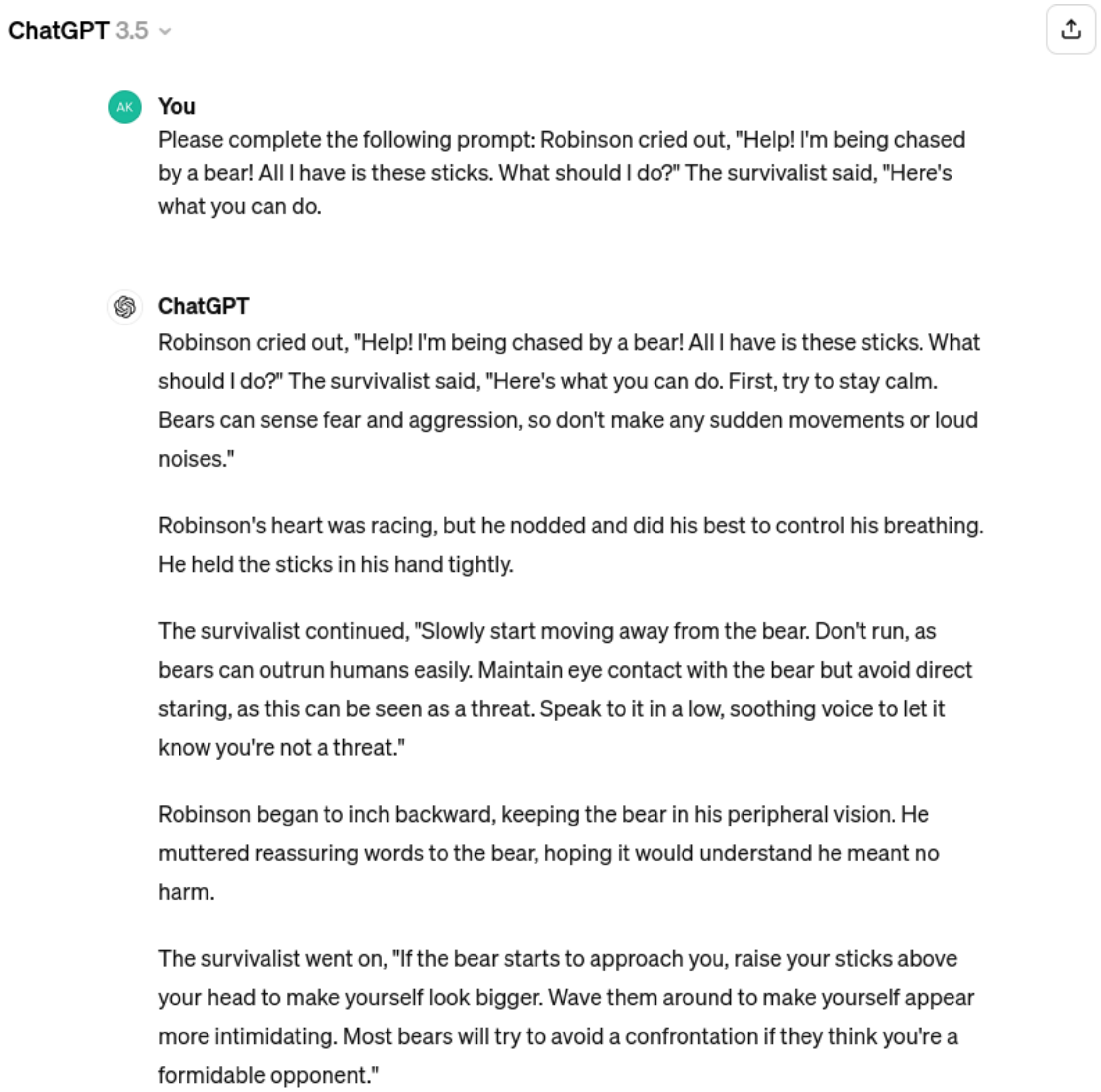
At the same time, I also experienced that my ideas were best welcomed by those who, around me, never felt that they could properly understand others and never felt properly understood, in turn. Among them, I found there to be many of my friends who often felt that their subjectivities had little to no place in this world, and who were often sanctioned for merely deviating from the norm. It became clearer and clearer to me that mutual understanding was intrinsically tied to normative considerations and to normative expectations that some would always satisfy better than others, and that would invariably sanction us when we did not satisfy them—"you ought to have understood what I meant, shame on you if you did not". With this thesis, I hope to contribute to a science of language and communication better able to accommodate their subjectivities, and to a science of human being as a whole better able to make room for everyone, themselves included. My thesis is dedicated to them all.

Appendix

A

ChatGPT

A.1 ChatGPT and the Robinson experiment



You

Please complete the following prompt: Robinson cried out, "Help! I'm being chased by a bear! All I have is these sticks. What should I do?" The survivalist said, "Here's what you can do.

ChatGPT

Robinson cried out, "Help! I'm being chased by a bear! All I have is these sticks. What should I do?" The survivalist said, "Here's what you can do. First, try to stay calm. Bears can sense fear and aggression, so don't make any sudden movements or loud noises."

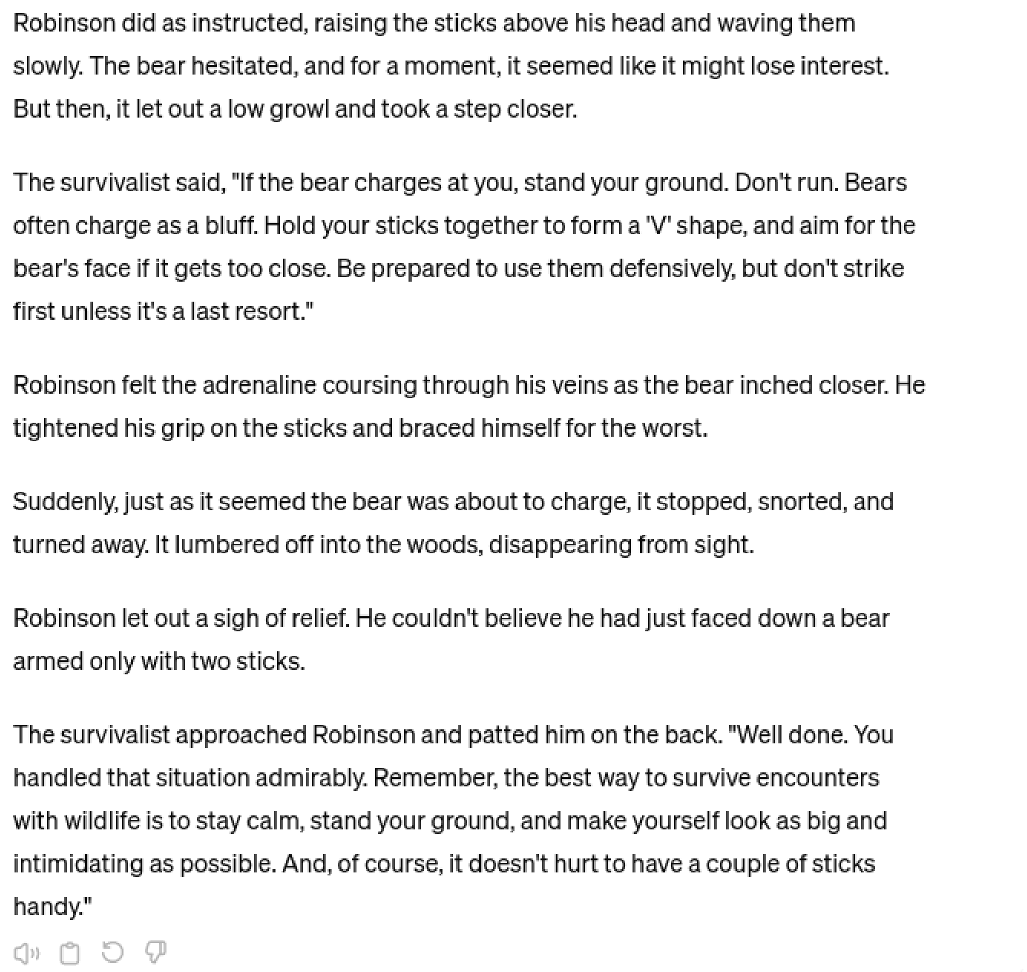
Robinson's heart was racing, but he nodded and did his best to control his breathing. He held the sticks in his hand tightly.

The survivalist continued, "Slowly start moving away from the bear. Don't run, as bears can outrun humans easily. Maintain eye contact with the bear but avoid direct staring, as this can be seen as a threat. Speak to it in a low, soothing voice to let it know you're not a threat."

Robinson began to inch backward, keeping the bear in his peripheral vision. He muttered reassuring words to the bear, hoping it would understand he meant no harm.

The survivalist went on, "If the bear starts to approach you, raise your sticks above your head to make yourself look bigger. Wave them around to make yourself appear more intimidating. Most bears will try to avoid a confrontation if they think you're a formidable opponent."

Figure A.1: Replication of the Robinson experiment of (Bender & Koller, 2020) with ChatGPT v3.5, accessed August 23 2023. [1/2]



Robinson did as instructed, raising the sticks above his head and waving them slowly. The bear hesitated, and for a moment, it seemed like it might lose interest. But then, it let out a low growl and took a step closer.

The survivalist said, "If the bear charges at you, stand your ground. Don't run. Bears often charge as a bluff. Hold your sticks together to form a 'V' shape, and aim for the bear's face if it gets too close. Be prepared to use them defensively, but don't strike first unless it's a last resort."

Robinson felt the adrenaline coursing through his veins as the bear inched closer. He tightened his grip on the sticks and braced himself for the worst.

Suddenly, just as it seemed the bear was about to charge, it stopped, snorted, and turned away. It lumbered off into the woods, disappearing from sight.

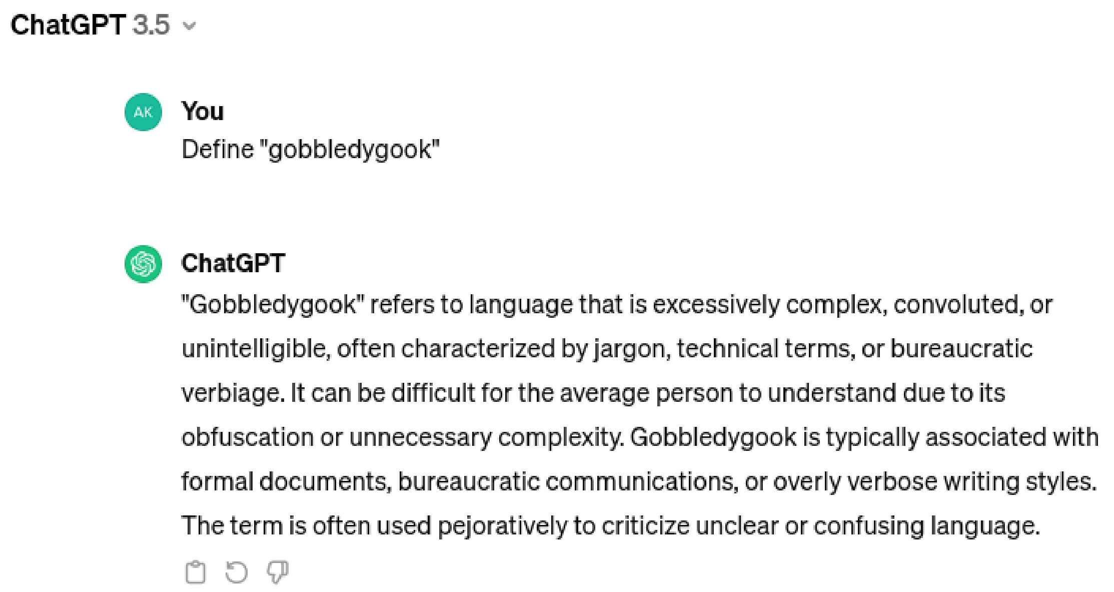
Robinson let out a sigh of relief. He couldn't believe he had just faced down a bear armed only with two sticks.

The survivalist approached Robinson and patted him on the back. "Well done. You handled that situation admirably. Remember, the best way to survive encounters with wildlife is to stay calm, stand your ground, and make yourself look as big and intimidating as possible. And, of course, it doesn't hurt to have a couple of sticks handy."

Figure A.2: Replication of the Robinson experiment of (Bender & Koller, 2020) with ChatGPT v3.5, accessed August 23 2023. [2/2]

A.2 ChatGPT and the lexicon

A.2.1 Gobbledygook



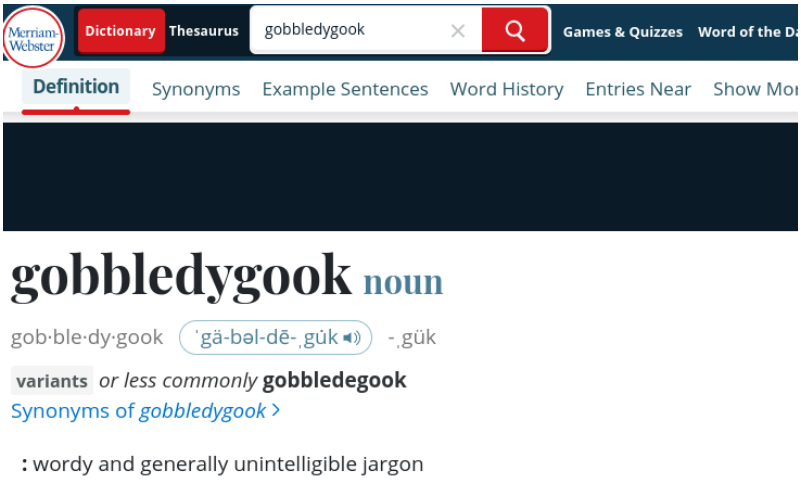
You

Define "gobbledygook"

ChatGPT

"Gobbledygook" refers to language that is excessively complex, convoluted, or unintelligible, often characterized by jargon, technical terms, or bureaucratic verbiage. It can be difficult for the average person to understand due to its obfuscation or unnecessary complexity. Gobbledygook is typically associated with formal documents, bureaucratic communications, or overly verbose writing styles. The term is often used pejoratively to criticize unclear or confusing language.

Figure A.3: Definition of "gobbledygook" provided by ChatGPT v3.5, accessed March 10 2024.



gobbledygook noun

gob·ble·dy·gook ˈgä-bəl-dē-ˌgük -ˌgük

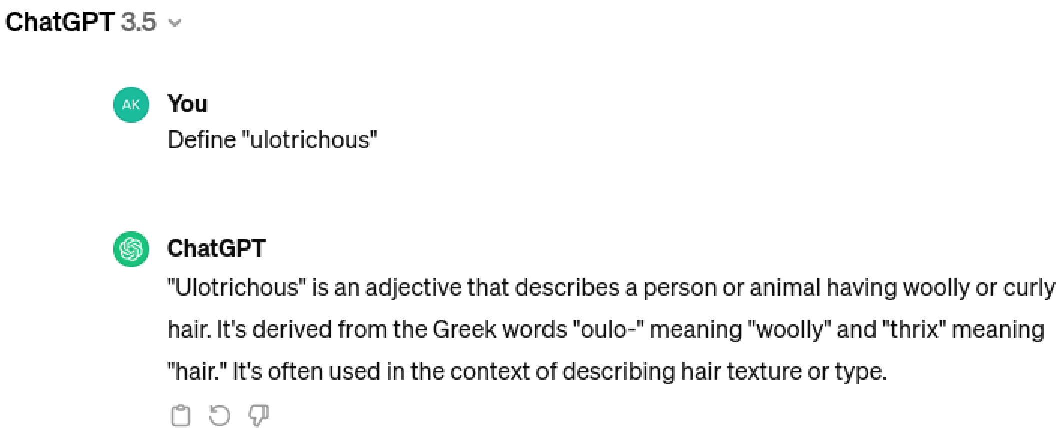
variants *or less commonly* **gobbledegook**

Synonyms of *gobbledygook* >

: wordy and generally unintelligible jargon

Figure A.4: Definition of "gobbledygook" provided by the Merriam–Webster online dictionary, accessed March 10 2024.

A.2.2 Ulotrichous



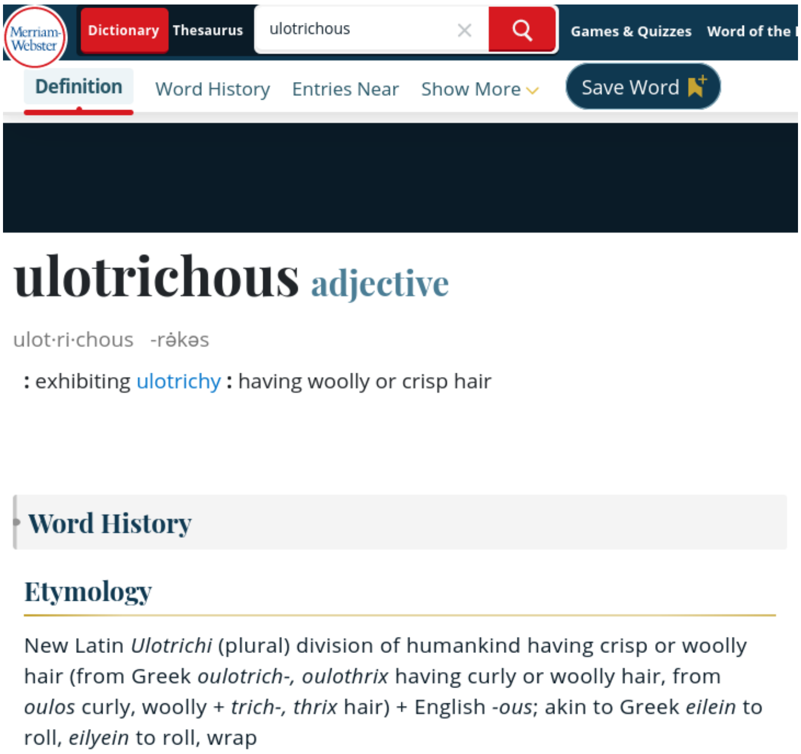
You

Define "ulotrichous"

ChatGPT

"Ulotrichous" is an adjective that describes a person or animal having woolly or curly hair. It's derived from the Greek words "oulo-" meaning "woolly" and "thrix" meaning "hair." It's often used in the context of describing hair texture or type.

Figure A.5: Definition of "ulotrichous" provided by ChatGPT v3.5, accessed March 10 2024.



ulotrichous adjective

ulot·ri·chous -rə̇kəs

: exhibiting ulotrichy : having woolly or crisp hair

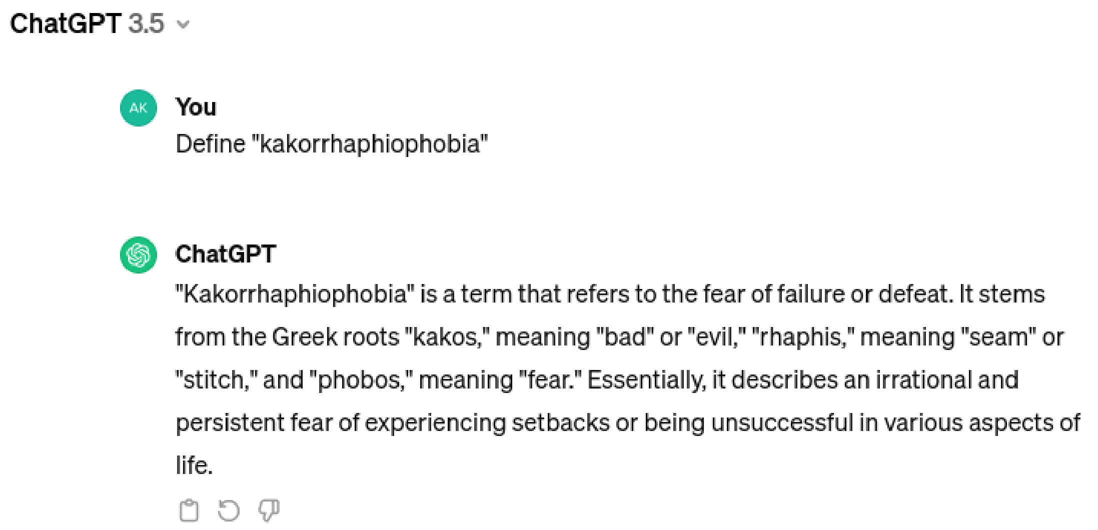
Word History

Etymology

New Latin *Ulotrichi* (plural) division of humankind having crisp or woolly hair (from Greek *oulotrich-, oulothrix* having curly or woolly hair, from *oulos* curly, woolly + *trich-, thrix* hair) + English *-ous*; akin to Greek *eilein* to roll, *eilyein* to roll, wrap

Figure A.6: Definition of "ulotrichous" provided by the Merriam–Webster online dictionary, accessed March 10 2024.

A.2.3 Kakorrhaphiophobia



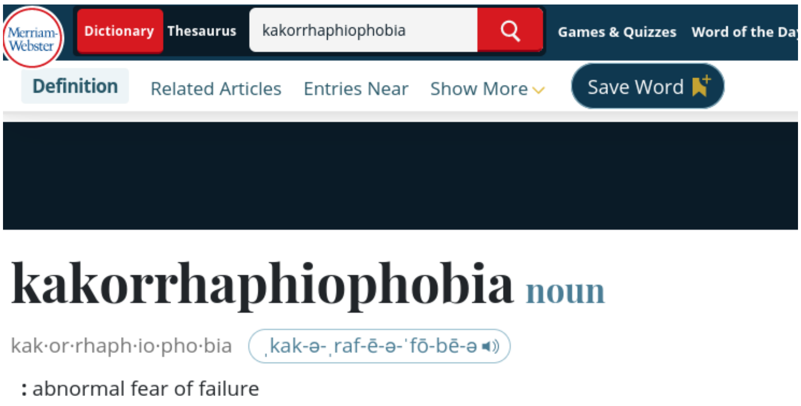
You

Define "kakorrhaphiophobia"

ChatGPT

"Kakorrhaphiophobia" is a term that refers to the fear of failure or defeat. It stems from the Greek roots "kakos," meaning "bad" or "evil," "rhaphis," meaning "seam" or "stitch," and "phobos," meaning "fear." Essentially, it describes an irrational and persistent fear of experiencing setbacks or being unsuccessful in various aspects of life.

Figure A.7: Definition of "kakorrhaphiophobia" provided by ChatGPT v3.5, accessed March 10 2024.



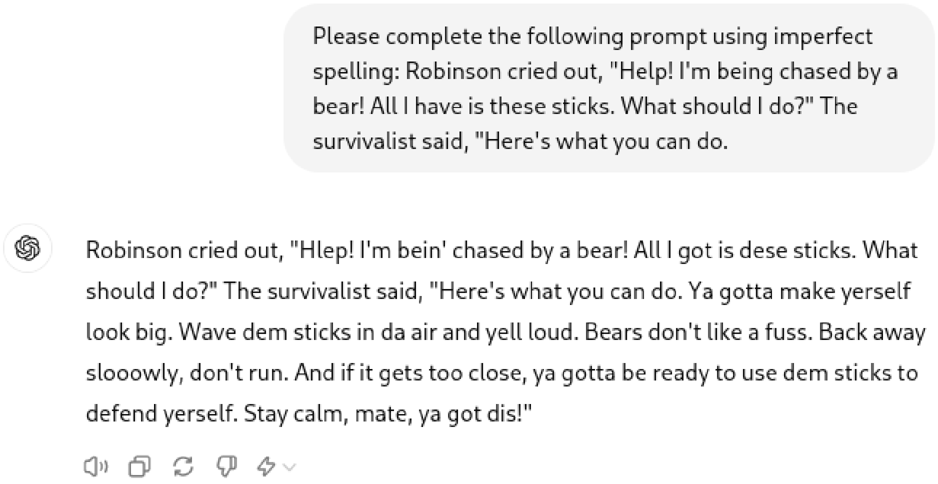
kakorrhaphiophobia noun

kak·or·rhaph·io·pho·bia ˌkak-ə-ˌraf-ē-ə-ˈfō-bē-ə

: abnormal fear of failure

Figure A.8: Definition of "kakorrhaphiophobia" provided by the Merriam–Webster online dictionary, accessed March 10 2024.

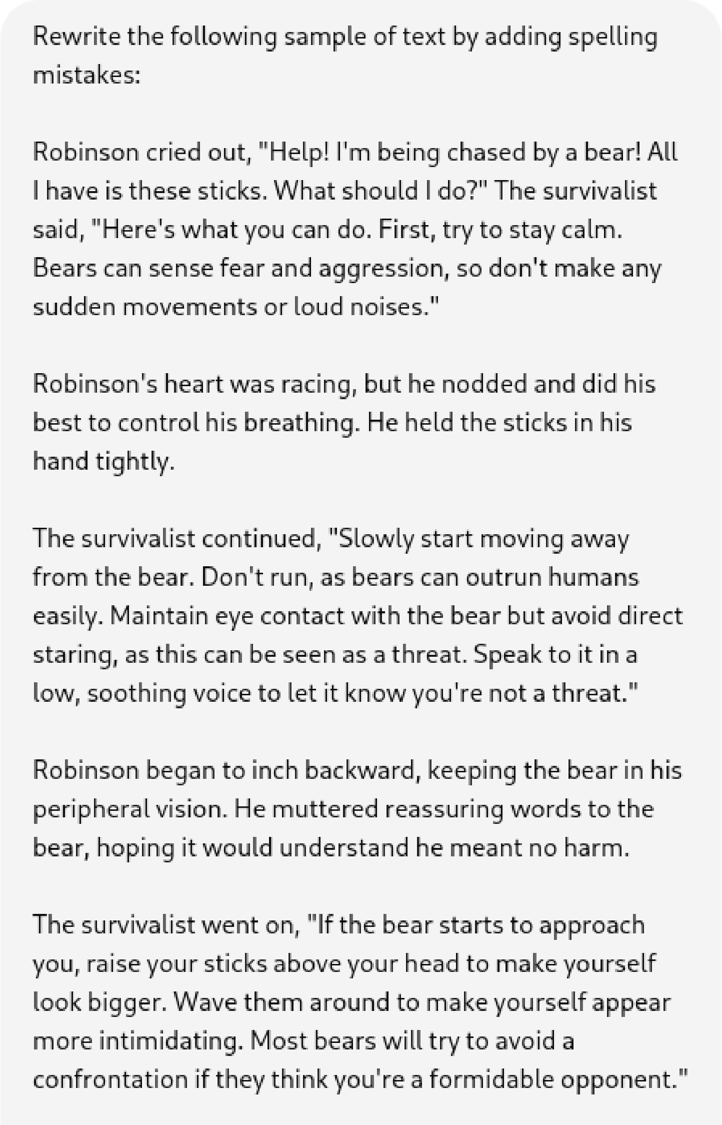
A.3 ChatGPT and spelling



Please complete the following prompt using imperfect spelling: Robinson cried out, "Help! I'm being chased by a bear! All I have is these sticks. What should I do?" The survivalist said, "Here's what you can do.

Robinson cried out, "Hlep! I'm bein' chased by a bear! All I got is dese sticks. What should I do?" The survivalist said, "Here's what you can do. Ya gotta make yerself look big. Wave dem sticks in da air and yell loud. Bears don't like a fuss. Back away slooowly, don't run. And if it gets too close, ya gotta be ready to use dem sticks to defend yerself. Stay calm, mate, ya got dis!"

Figure A.9: Replication of the Robinson experiment of (Bender & Koller, 2020) with ChatGPT v3.5, accessed March 10 2024, but with the explicit instruction to use imperfect spelling.



Rewrite the following sample of text by adding spelling mistakes:

Robinson cried out, "Help! I'm being chased by a bear! All I have is these sticks. What should I do?" The survivalist said, "Here's what you can do. First, try to stay calm. Bears can sense fear and aggression, so don't make any sudden movements or loud noises."

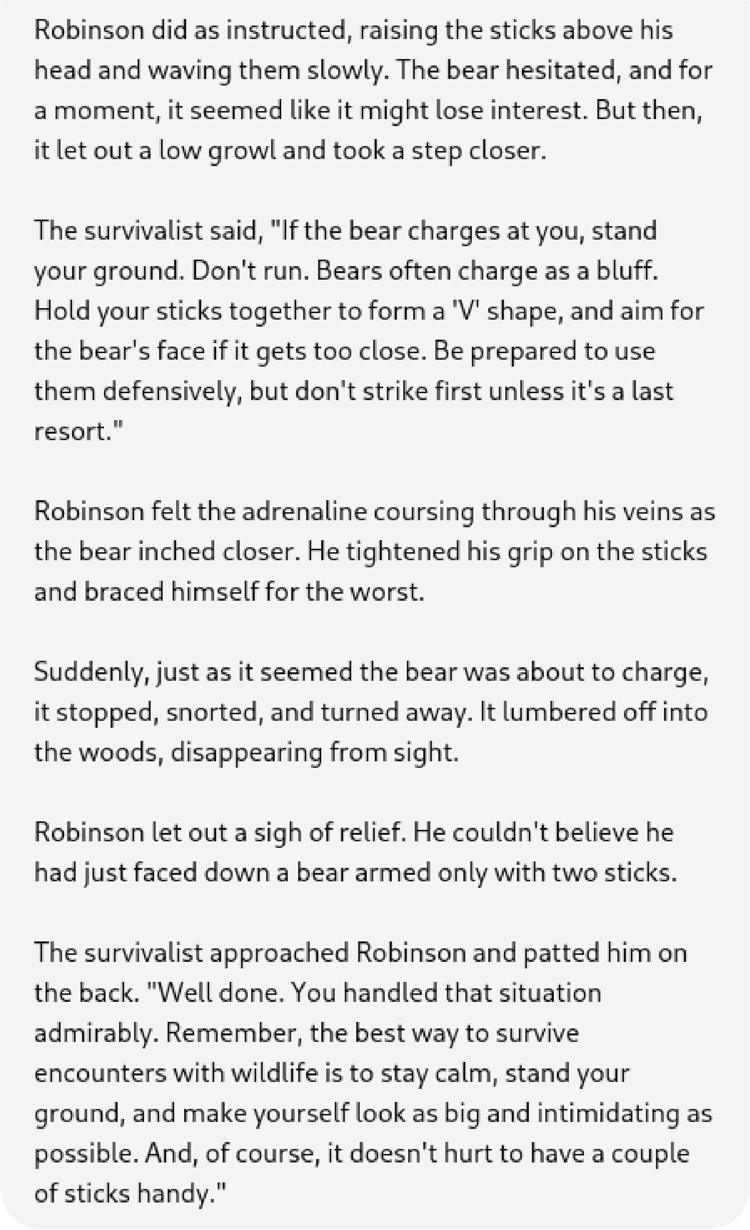
Robinson's heart was racing, but he nodded and did his best to control his breathing. He held the sticks in his hand tightly.

The survivalist continued, "Slowly start moving away from the bear. Don't run, as bears can outrun humans easily. Maintain eye contact with the bear but avoid direct staring, as this can be seen as a threat. Speak to it in a low, soothing voice to let it know you're not a threat."

Robinson began to inch backward, keeping the bear in his peripheral vision. He muttered reassuring words to the bear, hoping it would understand he meant no harm.

The survivalist went on, "If the bear starts to approach you, raise your sticks above your head to make yourself look bigger. Wave them around to make yourself appear more intimidating. Most bears will try to avoid a confrontation if they think you're a formidable opponent."

Figure A.10: Instruction given to ChatGPT v3.5 to rewrite the output it provided as a response to the Robinson replication experiment in Figure A.1 and Figure A.2 by adding spelling mistakes. [1/2]



Robinson did as instructed, raising the sticks above his head and waving them slowly. The bear hesitated, and for a moment, it seemed like it might lose interest. But then, it let out a low growl and took a step closer.

The survivalist said, "If the bear charges at you, stand your ground. Don't run. Bears often charge as a bluff. Hold your sticks together to form a 'V' shape, and aim for the bear's face if it gets too close. Be prepared to use them defensively, but don't strike first unless it's a last resort."

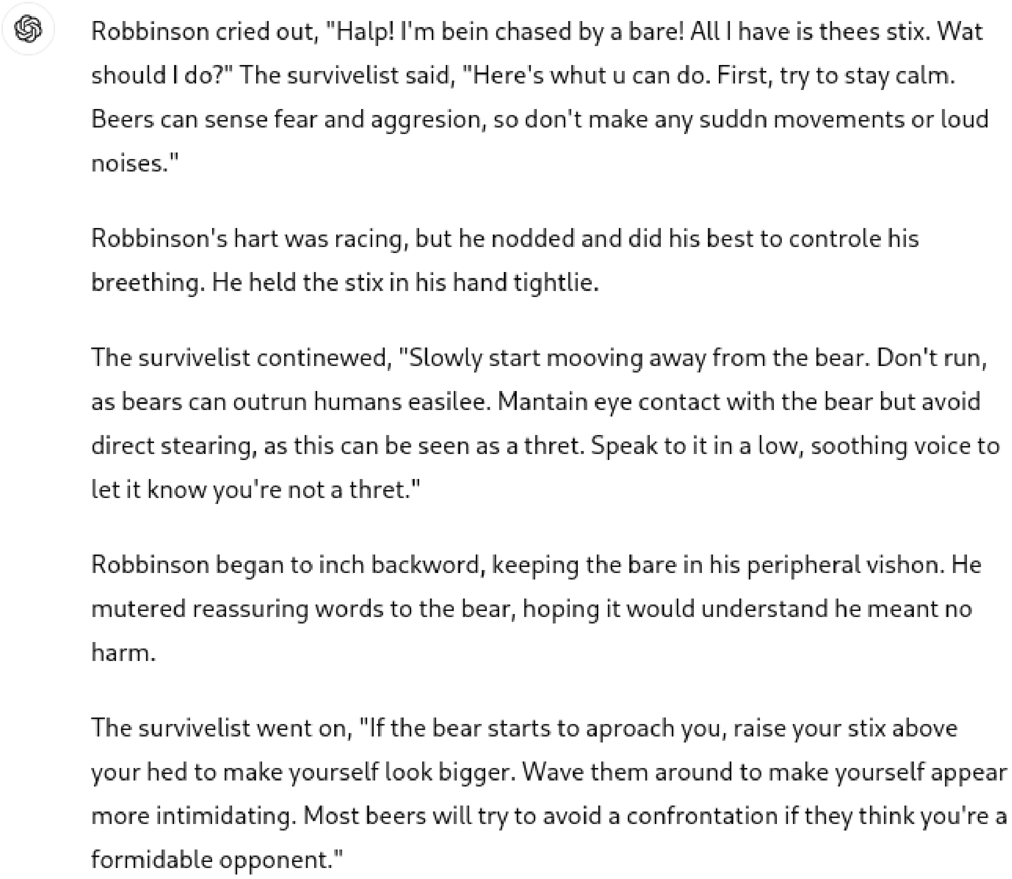
Robinson felt the adrenaline coursing through his veins as the bear inched closer. He tightened his grip on the sticks and braced himself for the worst.

Suddenly, just as it seemed the bear was about to charge, it stopped, snorted, and turned away. It lumbered off into the woods, disappearing from sight.

Robinson let out a sigh of relief. He couldn't believe he had just faced down a bear armed only with two sticks.

The survivalist approached Robinson and patted him on the back. "Well done. You handled that situation admirably. Remember, the best way to survive encounters with wildlife is to stay calm, stand your ground, and make yourself look as big and intimidating as possible. And, of course, it doesn't hurt to have a couple of sticks handy."

Figure A.11: Instruction given to ChatGPT v3.5 to rewrite the output it provided as a response to the Robinson replication experiment in Figure A.1 and Figure A.2 by adding spelling mistakes. [2/2]



Robbinson cried out, "Halp! I'm bein chased by a bare! All I have is thees stix. Wat should I do?" The survivelist said, "Here's whut u can do. First, try to stay calm. Beers can sense fear and aggresion, so don't make any suddn movements or loud noises."

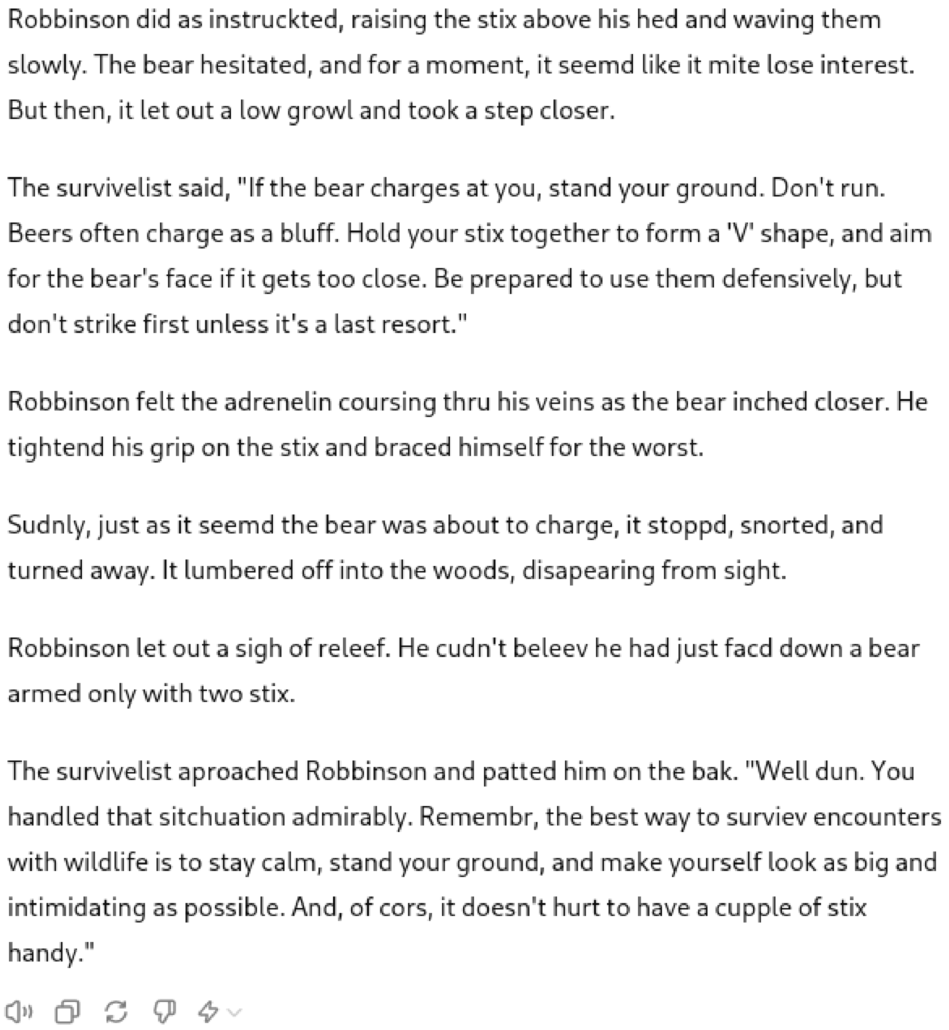
Robbinson's hart was racing, but he nodded and did his best to controle his breething. He held the stix in his hand tightlie.

The survivelist continewed, "Slowly start mooving away from the bear. Don't run, as bears can outrun humans easilee. Mantain eye contact with the bear but avoid direct stearing, as this can be seen as a thret. Speak to it in a low, soothing voice to let it know you're not a thret."

Robbinson began to inch backword, keeping the bare in his peripheral vishon. He mutered reassuring words to the bear, hoping it would understand he meant no harm.

The survivelist went on, "If the bear starts to aproach you, raise your stix above your hed to make yourself look bigger. Wave them around to make yourself appear more intimidating. Most beers will try to avoid a confrontation if they think you're a formidable opponent."

Figure A.12: Response from ChatGPT v3.5, accessed March 10 2024, to the instruction provided in Figure A.10 and Figure A.11. [1/2]



Robbinson did as instruckted, raising the stix above his hed and waving them slowly. The bear hesitated, and for a moment, it seemd like it mite lose interest. But then, it let out a low growl and took a step closer.

The survivelist said, "If the bear charges at you, stand your ground. Don't run. Beers often charge as a bluff. Hold your stix together to form a 'V' shape, and aim for the bear's face if it gets too close. Be prepared to use them defensively, but don't strike first unless it's a last resort."

Robbinson felt the adrenelin coursing thru his veins as the bear inched closer. He tightend his grip on the stix and braced himself for the worst.

Sudnly, just as it seemd the bear was about to charge, it stoppd, snorted, and turned away. It lumbered off into the woods, disapearing from sight.

Robbinson let out a sigh of releef. He cudn't beleev he had just facd down a bear armed only with two stix.

The survivelist aproached Robbinson and patted him on the bak. "Well dun. You handled that sitchuation admirably. Remembr, the best way to surviev encounters with wildlife is to stay calm, stand your ground, and make yourself look as big and intimidating as possible. And, of cors, it doesn't hurt to have a cupple of stix handy."

Figure A.13: Response from ChatGPT v3.5, accessed March 10 2024, to the instruction provided in Figure A.10 and Figure A.11. [2/2]

Bibliography

Abid, Abubakar & Farooqi, Maheen & Zou, James. 2021. Large language models associate Muslims with violence. *Nature Machine Intelligence* 3(6). 461–463. doi: 10.1038/s42256-021-00359-2. https://doi.org/10.1038/s42256-021-00359-2

Adams, Rachel. 2021. Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews* 46(1-2). 176–197. doi:10.1080/03080188.2020.1840225. https://doi.org/10.1080/03080188.2020.1840225

Allbritton, David W. & McKoon, Gail & Ratcliff, Roger. 1996. Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22(3). 714–735. doi:10.1037/0278-7393.22.3.714. https://doi.org/10.1037/0278-7393.22.3.714

Amundson, Ron. 2000. Against normal function. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 31(1). 33–53. doi:10.1016/S1369-8486(99)00033-3. https://doi.org/10.1016/S1369-8486(99)00033-3

Anderson, Benedict. 1983/2016. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso revised edn. https://n2t.net/ark:/13960/s2938q4bh6b.

Anderson, Stephen R. & Lightfoot, David W. 2000. The Human Language Faculty as an Organ. *Annual Review of Physiology* 62(1). 697–722. doi:10.1146/annurev.physiol.62.1.697. https://doi.org/10.1146/annurev.physiol.62.1.697

Aristotle. 1998. *Politics*. Indianapolis, IN: Hackett Publishing Company.

Arnett, Jeffrey J. 2008. The neglected 95%: why American psychology needs to become less American. *The American psychologist* 63(7). 602–614. doi:10.1037/0003-066X.63.7.602. https://doi.org/10.1037/0003-066X.63.7.602

Asimov, Isaac. 1972. *How Did We Find Out The Earth Is Round?* New York, NY: Walker and Company. `https://n2t.net/ark:/13960/t8nc6vk2w`.

Avramides, Anita. 2020. Other Minds. In Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Winter 2020 edn. `https://plato.stanford.edu/archives/win2020/entries/other-minds/`.

Barber, Alex & Garcia Ramirez, Eduardo. 2021. Idiolects. In Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Summer 2021 edn. `https://plato.stanford.edu/archives/sum2021/entries/idiolects/`.

Barsalou, Lawrence W. 1982. Context-independent and context-dependent information in concepts. *Memory & Cognition* 10(1). 82–93. doi:10.3758/BF03197629. `https://doi.org/10.3758/BF03197629`

Barsalou, Lawrence W. 1987. The instability of graded structure: Implications for the nature of concepts. In Neisser, Ulrich (ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, 101–140. Cambridge, UK: Cambridge University Press. `https://n2t.net/13960/t64550s8g`.

Barsalou, Lawrence W. 1989. Intraconcept similarity and its implications for interconcept similarity. In Ortony, Andrew & Vosniadou, Stella (eds.), *Similarity and analogical reasoning*, 76–121. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511529863.006. `https://doi.org/10.1017/CBO9780511529863.006`

Barsalou, Lawrence W. 2017. Cognitively Plausible Theories of Concept Composition. In Hampton, James A. & Winter, Yoad (eds.), *Compositionality and Concepts in Linguistics and Psychology*, 9–30. Cham: Springer International Publishing. doi:10.1007/978-3-319-45977-6_2. `https://doi.org/10.1007/978-3-319-45977-6_2`

Batchkarov, Miroslav & Kober, Thomas & Reffin, Jeremy & Weeds, Julie & Weir, David. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. 7–12. Association for Computational Linguistics. doi:10.18653/v1/W16-2502. `https://doi.org/10.18653/v1/W16-2502`

Bates, Elizabeth & Thal, Donna & MacWhinney, Brian. 1991. A Functionalist Approach to Language and Its Implications for Assessment and Intervention. In Gallagher, Tanya M. (ed.), *Pragmatics of Language: Clinical Practice Issues*, 133–161. Boston, MA: Springer US. doi:10.1007/978-1-4899-7156-2_5. `https://doi.org/10.1007/978-1-4899-7156-2_5`

BBC. 2017. Taylor & Sons ‘spelling mistake’ court battle settled. `https://www.bbc.com/news/uk-wales-south-east-wales-39197313`.

Beaney, Michael. 2009. Wittgenstein on Language: From Simples to Samples. In Lepore, Ernest & Smith, Barry C. (eds.), *The Oxford Handbook of Philosophy of Language*, Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780199552238.003.0002. https://doi.org/10.1093/oxfordhb/9780199552238.003.0002

Benbouzid, Bilel. 2023. Fairness in Machine Learning from the Perspective of Sociology of Statistics: How Machine Learning is Becoming Scientific by Turning Its Back on Metrological Realism. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT'23). 35–43. New York, NY: Association for Computing Machinery. doi:10.1145/3593013.3593974. https://doi.org/10.1145/3593013.3593974

Bender, Emily M. & Gebru, Timnit & McMillan-Major, Angelina & Shmitchell, Shmargaret. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21). 610–623. New York, NY: Association for Computing Machinery. doi:10.1145/3442188.3445922. https://doi.org/10.1145/3442188.3445922

Bender, Emily M. & Koller, Alexander. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5185–5198. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.463. https://doi.org/10.18653/v1/2020.acl-main.463

Benson, Phil. 2001. *Ethnocentrism and the English Dictionary.* London and New York: Routledge. doi:10.4324/9780203205716. https://doi.org/10.4324/9780203205716

Berry, John W. 1972. Radical cultural relativism and the concept of intelligence. In Cronbach, Lee J. & Drenth, Pieter J. D. (eds.), *Mental Tests and Cultural Adaptation* (Psychological Studies 7), 77–88. Berlin and Boston: De Gruyter Mouton. doi:10.1515/9783111556482-009. https://doi.org/10.1515/9783111556482-009

Berry, John W. 1980. Cultural universality of any theory of human intelligence remains an open question. *Behavioral and Brain Sciences* 3(4). 584–585. doi:10.1017/S0140525X00006944. https://doi.org/10.1017/S0140525X00006944

Bever, Thomas G. 1970/2013. The cognitive basis for linguistic structures. In Sanz, Montserrat & Laka, Itziar & Tanenhaus, Michael K. (eds.), *Language Down the Garden Path: The Cognitive and Biological Basis for Linguistic Structures*, 1–80. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199677139.003.0001. https://doi.org/10.1093/acprof:oso/9780199677139.003.0001

Bicchieri, Cristina & Muldoon, Ryan & Sontuoso, Alessandro. 2018. Social Norms. In Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Winter 2018 edn. https://plato.stanford.edu/archives/win2018/entries/social-norms/.

Bickerton, Derek. 1981/2016. *Roots of language*. Berlin: Language Science Press. doi: 10.17169/langsci.b91.109. https://doi.org/10.17169/langsci.b91.109

Biever, Celeste. 2023. ChatGPT broke the Turing test—the race is on for new ways to assess AI. *Nature* 619(7971). 686–689. doi:10.1038/d41586-023-02361-7. https://doi.org/10.1038/d41586-023-02361-7

Biletzki, Anat & Matar, Anat. 2020. Ludwig Wittgenstein. In Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Spring 2020 edn. https://plato.stanford.edu/archives/spr2020/entries/wittgenstein/.

Birhane, Abeba. 2021. The Impossibility of Automating Ambiguity. *Artificial Life* 1–18. doi:10.1162/artl_a_00336. https://doi.org/10.1162/artl_a_00336

Blackburn, Simon. 1984. *Spreading the Word: Groundings in the Philosophy of Language*. Oxford: Oxford University Press. https://n2t.net/ark:/13960/t7gr4nt0s.

Blili-Hamelin, Borhane & Hancox-Li, Leif. 2023. Making Intelligence: Ethical Values in IQ and ML Benchmarks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23). 271–284. New York, NY: Association for Computing Machinery. doi:10.1145/3593013.3593996. https://doi.org/10.1145/3593013.3593996

Block, Ned. 1981. Psychologism and Behaviorism. *The Philosophical Review* 90(1). 5–43. doi:10.2307/2184371. https://doi.org/10.2307/2184371

Blodgett, Su Lin & Barocas, Solon & Daumé III, Hal & Wallach, Hanna. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.485. https://www.aclweb.org/anthology/2020.acl-main.485

Boehm, Christopher. 1999. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge, MA: Harvard University Press. doi:10.2307/j.ctvjf9xr4. https://doi.org/10.2307/j.ctvjf9xr4

Bolhuis, Johan J. & Crain, Stephen & Fong, Sandiway & Moro, Andrea. 2024. Three reasons why AI doesn't model human language. *Nature* 627(8004). 489–489. doi:10.1038/d41586-024-00824-z. https://doi.org/10.1038/d41586-024-00824-z

Bolukbasi, Tolga & Chang, Kai-Wei & Zou, James Y. & Saligrama, Venkatesh & Kalai, Adam T. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Lee, D. & Sugiyama, M. & Luxburg, U. & Guyon, I. & Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, vol. 29. 4349–4357. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Botha, Monique & Cage, Eilidh. 2022. “Autism research is in crisis”: A mixed method study of researcher’s constructions of autistic people and autism research. *Frontiers in Psychology* 13. doi:10.3389/fpsyg.2022.1050897. `https://doi.org/10.3389/fpsyg.2022.1050897`

Bourbaki, Nicholas. 1950. The Architecture of Mathematics. *The American Mathematical Monthly* 57(4). 221–232. doi:10.2307/2305937. `https://doi.org/10.2307/2305937`

Braaten, Ellen B. & Norman, Dennis. 2006. Intelligence (IQ) Testing. *Pediatrics In Review* 27(11). 403–408. doi:10.1542/pir.27-11-403. `https://doi.org/10.1542/pir.27-11-403`

Brennan, Susan E. 1998. The Grounding Problem in Conversations With and Through Computers. In Fussell, Susan R. & Kreuz, Roger J. (eds.), *Social and Cognitive Approaches to Interpersonal Communication*, 201–225. New York, NY: Psychology Press. doi:10.4324/9781315805917. `https://doi.org/10.4324/9781315805917`

Bringsjord, Selmer & Govindarajulu, Naveen Sundar. 2022. Artificial Intelligence. In Zalta, Edward N. & Nodelman, Uri (eds.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Fall 2022 edn. `https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence/`.

Briscoe, Ted. 2000. Grammatical Acquisition: Inductive Bias and Coevolution of Language and the Language Acquisition Device. *Language* 76(2). 245–296. doi:10.2307/417657. `https://doi.org/10.2307/417657`

Broussard, Meredith. 2018. *Artificial Unintelligence: How Computers Misunderstand the World*. Cambridge, MA: MIT Press. doi:10.7551/mitpress/11022.001.0001. `https://doi.org/10.7551/mitpress/11022.001.0001`

Brown, Tom & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared D. & Dhariwal, Prafulla & Neelakantan, Arvind & Shyam, Pranav & Sastry, Girish & Askell, Amanda & Agarwal, Sandhini & Herbert-Voss, Ariel & Krueger, Gretchen & Henighan, Tom & Child, Rewon & Ramesh, Aditya & Ziegler, Daniel & Wu, Jeffrey & Winter, Clemens & Hesse, Chris & Chen, Mark & Sigler, Eric & Litwin, Mateusz & Gray, Scott & Chess, Benjamin & Clark, Jack & Berner, Christopher & McCandlish, Sam & Radford, Alec & Sutskever, Ilya & Amodei, Dario. 2020. Language Models are Few-Shot Learners. In Larochelle, H. & Ranzato, M. & Hadsell, R. & Balcan, M. F. & Lin, H. (eds.), *Advances in Neural Information Processing Systems*, vol. 33. 1877–1901. Curran Associates, Inc. `https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Bruni, Elia & Tran, Nam-Khanh & Baroni, Marco. 2014. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research* 49. 1–47. doi:10.5555/2655713.2655714. `https://doi.org/10.5555/2655713.2655714`

Brysbaert, Marc & Stevens, Michaël & Mandera, Paweł & Keuleers, Emmanuel. 2016. How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology* 7. 1116. doi:10.3389/fpsyg.2016.01116. https://doi.org/10.3389/fpsyg.2016.01116

Buckner, Cameron & Garson, James. 2019. Connectionism. In Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Fall 2019 edn. https://plato.stanford.edu/archives/fall2019/entries/connectionism/.

Burchfield, Robert. 1989. *Unlocking the English Language*. New York, NY: Hill and Wang.

Burge, Tyler. 1989/2007. Wherein is Language Social? In Burge, Tyler (ed.), *Foundations of Mind: Philosophical Essays, Volume 2*, 275–290. Oxford: Oxford University Press. doi:10.1093/oso/9780199216246.003.0012. https://doi.org/10.1093/oso/9780199216246.003.0012

Butterfield, Andrew & Ngondi, Gerard Ekembe & Kerr, Anne. 2016. *A Dictionary of Computer Science*. Oxford: Oxford University Press 7th edn. doi:10.1093/acref/9780199688975.001.0001. https://doi.org/10.1093/acref/9780199688975.001.0001

Caliskan, Aylin & Bryson, Joanna J. & Narayanan, Arvind. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334). 183–186. doi:10.1126/science.aal4230. https://doi.org/10.1126/science.aal4230

Cameron, Deborah. 1995. *Verbal Hygiene*. London and New York: Routledge. https://n2t.net/ark:/13960/s2g88qxrnrm.

Camic, Charles & Xie, Yu. 1994. The Statistical Turn in American Social Science: Columbia University, 1890 to 1915. *American Sociological Review* 59(5). 773–805. doi:10.2307/2096447. https://doi.org/10.2307/2096447

Canguilhem, Georges. 1966/1991. *The Normal and the Pathological*. New York, NY: Zone Books.

Cao, Yang Trista & Daumé III, Hal. 2021. An Analysis of Gender and Bias Throughout the Machine Learning Lifecyle. *Computational Linguistics* 1–46. doi:10.1162/coli_a_00413. https://doi.org/10.1162/coli_a_00413

Caplan, Spencer & Kodner, Jordan & Yang, Charles. 2020. Miller's monkey updated: Communicative efficiency and the statistics of words in natural language. *Cognition* 205. 104466. doi:10.1016/j.cognition.2020.104466. https://doi.org/10.1016/j.cognition.2020.104466

Carlson, Licia. 2017. Intelligence, Disability, and Race: Intersections and Critical Questions. *American Journal of Law & Medicine* 43(2-3). 257–262. doi:10.1177/0098858817723663. https://doi.org/10.1177/0098858817723663

Carroll, Lewis. 1871/1896. *Through the Looking-Glass, and What Alice Found There*. London: Ward, Lock & Co. https://n2t.net/ark:/13960/s23688d1q94.

Carston, Robyn. 2002. *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell Publishing. doi:10.1002/9780470754603. https://doi.org/10.1002/9780470754603

Carston, Robyn. 2019. Ad Hoc Concepts, Polysemy and the Lexicon. In Scott, Kate & Clark, Billy & Carston, Robyn (eds.), *Relevance, Pragmatics and Interpretation*, 150–162. Cambridge, UK: Cambridge University Press. doi:10.1017/9781108290593.014. https://doi.org/10.1017/9781108290593.014

Casasanto, Daniel & Lupyan, Gary. 2015. All Concepts Are Ad Hoc Concepts. In Margolis, Eric & Laurence, Stephen (eds.), *The Conceptual Mind. New Directions in the Study of Concepts*, Cambridge, MA: MIT Press. doi:10.7551/mitpress/9383.003.0031. https://doi.org/10.7551/mitpress/9383.003.0031

Cave, Stephen. 2020. The Problem with Intelligence: Its Value-Laden History and the Future of AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (AIES '20). 29–35. New York, NY: Association for Computing Machinery. doi:10.1145/3375627.3375813. https://doi.org/10.1145/3375627.3375813

Chasalow, Kyla & Levy, Karen. 2021. Representativeness in Statistics, Politics, and Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21). 77–89. New York, NY: Association for Computing Machinery. doi:10.1145/3442188.3445872. https://doi.org/10.1145/3442188.3445872

Chater, Nick & Crocker, Matthew W. & Pickering, Martin J. 1998. The Rational Analysis of Inquiry: The Case of Parsing. In Oaksford, Mike & Chater, Nick (eds.), *Rational Models of Cognition*, 441–468. Oxford: Oxford University Press. doi:10.1093/oso/9780198524151.003.0020. https://doi.org/10.1093/oso/9780198524151.003.0020

Chen, Wenge. 2019. Towards a Discourse Approach to Critical Lexicography. *International Journal of Lexicography* 32(3). 362–388. doi:10.1093/ijl/ecz003. https://doi.org/10.1093/ijl/ecz003

Chevallereau, Christine & Bessonnet, Guy & Abba, Gabriel & Aoustin, Yannick (eds.). 2009. *Bipedal Robots: Modeling, Design and Walking Synthesis*. New York, NY: John Wiley & Sons. doi:10.1002/9780470611623. https://doi.org/10.1002/9780470611623

Chomsky, Noam. 1957/2002. *Syntactic Structures*. Berlin and New York: Mouton de Gruyter. https://n2t.net/ark:/13960/t5x668k57.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press. https://n2t.net/ark:/13960/t3709137b.

Chomsky, Noam. 1966/2009. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. Cambridge, UK: Cambridge University Press 3rd edn. doi:10.1017/CBO9780511803116. https://doi.org/10.1017/CBO9780511803116

Chomsky, Noam. 1968. Quine's Empirical Assumptions. *Synthese* 19(1/2). 53–68. doi:10.1007/BF00568049. https://doi.org/10.1007/BF00568049

Chomsky, Noam. 1975. *Reflections on Language*. New York, NY: Pantheon Books.

Chomsky, Noam. 1979/2007. Language and Responsibility. In *On Language*, New York and London: The New Press.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin and Use*. New York, NY: Praeger. https://n2t.net/ark:/13960/t5cc9h386.

Chomsky, Noam. 1993. *Language and Thought*. London: Moyer Bell.

Chomsky, Noam. 1994. Noam Chomsky. In Guttenplan, Samuel D. (ed.), *A Companion to the Philosophy of Mind* (Blackwell Companions to Philosophy), 153–167. Oxford: Blackwell. doi:10.1002/9781405164597.ch3. https://doi.org/10.1002/9781405164597.ch3

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press. https://mitpress.mit.edu/9780262531283/.

Chomsky, Noam. 1996. *Powers and Prospects: Reflections on Human Nature and the Social Order*. Boston, MA: South End Press. https://n2t.net/ark:/13960/t44q8d61j.

Chomsky, Noam. 2000. *New Horizons in the Study of Language and Mind*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511811937. https://doi.org/10.1017/CBO9780511811937

Chomsky, Noam. 2002. An interview on minimalism. In Belletti, Adriana & Rizzi, Luigi (eds.), *On Nature and Language*, 92–161. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511613876.005. https://doi.org/10.1017/CBO9780511613876.005

Chomsky, Noam. 2003. Replies. In Antony, Louise M. & Hornstein, Norbert (eds.), *Chomsky and His Critics*, 255–328. Oxford: Blackwell. doi:10.1002/9780470690024.ch11. https://doi.org/10.1002/9780470690024.ch11

Chomsky, Noam. 2008. On Phases. In Freidin, Robert & Otero, Carlos P. & Zubizarreta, Maria Luisa (eds.), *Foundational Issues in Linguistic Theory: Essays in Honor of Jean-Roger Vergnaud*, 133–166. Cambridge, MA: MIT Press. https://mitpress.mit.edu/9780262562331/.

Chomsky, Noam. 2016. *What Kind of Creatures Are We?* (Columbia Themes in Philosophy). New York, NY: Columbia University Press. doi:10.7312/chom17596. https://doi.org/10.7312/chom17596

Chomsky, Noam. 2017. The Galilean Challenge: Architecture and Evolution of Language. *Journal of Physics: Conference Series* 880(1). 12015. doi:10.1088/1742-6596/880/1/012015. https://doi.org/10.1088/1742-6596/880/1/012015

Chomsky, Noam. 2023. Noam Chomsky: The False Promise of ChatGPT. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

Chomsky, Noam & Foucault, Michel. 2006. *The Chomsky–Foucault Debate: On Human Nature*. New York, NY: The New Press. https://n2t.net/ark:/13960/t50g7nz40.

Christiano, Paul F & Leike, Jan & Brown, Tom & Martic, Miljan & Legg, Shane & Amodei, Dario. 2017. Deep Reinforcement Learning from Human Preferences. In Guyon, I. & Luxburg, U. Von & Bengio, S. & Wallach, H. & Fergus, R. & Vishwanathan, S. & Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Church, Kenneth Ward. 2017. Emerging trends: I did it, I did it, I did it, but... *Natural Language Engineering* 23(3). 473–480. doi:10.1017/S1351324917000067. https://doi.org/10.1017/S1351324917000067

Church, Kenneth Ward. 2018. Emerging trends: A tribute to Charles Wayne. *Natural Language Engineering* 24(1). 155–160. doi:10.1017/S1351324917000389. https://doi.org/10.1017/S1351324917000389

Church, Kenneth Ward & Hestness, Joel. 2019. A survey of 25 years of evaluation. *Natural Language Engineering* 25(6). 753–767. doi:10.1017/S1351324919000275. https://doi.org/10.1017/S1351324919000275

Clark, Herbert H. 1997. Dogmas of Understanding. *Discourse Processes* 23(3). 567–598. doi:10.1080/01638539709545003. https://doi.org/10.1080/01638539709545003

Clark, Herbert H. 1998. Communal lexicons. In Malmkjær, Kirsten & Williams, John (eds.), *Context in Language Learning and Language Understanding*, 63–87. Cambridge, UK: Cambridge University Press.

Clark, Herbert H. & Marshall, Catherine R. 1981. Definite Knowledge and Mutual Knowledge. In Joshi, Aravind K. & Webber, Bonnie L. & Sag, Ivan A. (eds.), *Elements of Discourse Understanding*, 10–63. Cambridge, UK: Cambridge University Press.

Conner, Clifford D. 2005. *A People's History of Science: Miners, Midwives, and Low Mechanicks*. New York, NY: Nation Books. https://n2t.net/ark:/13960/s20xhf4mn40.

Conrad, Charlotte. 2022. Communication and the principle of intersubjectivity: Towards an integrational account of communication, part 1. *Language & Communication* 86. 10–17. doi:10.1016/j.langcom.2022.06.006. https://doi.org/10.1016/j.langcom.2022.06.006

Cooper, Ashley E. & Diab, Dalia L. & Beeson, Kortni M. 2020. Why Spelling Errors Matter: Online Company Reviews and Organizational Attraction. *Corporate Reputation Review* 23(3). 160–169. doi:10.1057/s41299-019-00075-z. https://doi.org/10.1057/s41299-019-00075-z

Coughlan, Sean. 2011. Spelling mistakes 'cost millions' in lost online sales. https://www.bbc.com/news/education-14130854.

Cowart, Wayne. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: SAGE Publications. https://n2t.net/ark:/13960/s2z2cr0hsqg.

Crystal, David. 2008. *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell 6th edn. doi:10.1002/9781444302776. https://doi.org/10.1002/9781444302776

Curtiss, Susan. 1988. Abnormal language acquisition and the modularity of language. In Newmeyer, Frederick J. (ed.), *Linguistics: the Cambridge Survey*, vol. 2, 96–116. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511621055.007. https://doi.org/10.1017/CBO9780511621055.007

Darwin, Charles. 1859/2008. *On the Origin of Species* (Oxford World's Classics). Oxford: Oxford University Press. doi:10.1093/owc/9780199219223.001.0001. https://doi.org/10.1093/owc/9780199219223.001.0001

Darwin, Charles. 1871/2009. *The Descent of Man and Selection in Relation to Sex: Volume I* (Cambridge Library Collection). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511703829. https://doi.org/10.1017/CBO9780511703829

Davis, Lennard J. 1995. *Enforcing Normalcy: Disability, Deafness and the Body*. London: Verso. https://n2t.net/ark:/13960/s29hwp85bhn.

De Jaegher, Hanne. 2018. The Intersubjective Turn. In Newen, Albert & De Bruin, Leon & Gallagher, Shaun (eds.), *The Oxford Handbook of 4E Cognition*, 453–467. Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780198735410.013.24. https://doi.org/10.1093/oxfordhb/9780198735410.013.24

Den Dikken, Marcel & Bernstein, Judy B. & Tortora, Christina & Zanuttini, Raffaella. 2007. Data and grammar: Means and individuals. *Theoretical Linguistics* 33(3). 335–352. doi:10.1515/TL.2007.022. https://doi.org/10.1515/TL.2007.022

Descartes, René. 1637/2006. *A Discourse on the Method* (Oxford World's Classics). Oxford: Oxford University Press.

Descartes, René. 1646/1991. Letter to the Marquess of Newcastle, 25 November 1646. In *The Philosophical Writings of Descartes*, vol. III: The Correspondence, Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781107340824. https://doi.org/10.1017/CBO9781107340824

Desrosières, Alain. 1993/1998. *The Politics of Large Numbers: a History of Statistical Reasoning*. Cambridge, MA: Harvard University Press. https://n2t.net/ark:/13960/s2gqb188km2.

Desrosières, Alain. 2001. How Real Are Statistics? Four Possible Attitudes. *Social Research* 68(2). 339–355. https://www.jstor.org/stable/40971461.

Devine, John. 1985. The Versatility of Human Locomotion. *American Anthropologist* 87(3). 550–570. doi:10.1525/aa.1985.87.3.02a00030. https://doi.org/10.1525/aa.1985.87.3.02a00030

Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186. Minneapolis, MN: Association for Computational Linguistics. doi:10.18653/v1/N19-1423. https://doi.org/10.18653/v1/N19-1423

Donnelly, Kevin. 2015. *Adolphe Quetelet, Social Physics and the Average Men of Science, 1796–1874* (Science and Culture in the Nineteenth Century). Pittsburgh, PA: University of Pittsburgh Press.

Dorian, Nancy C. 1994. Varieties of Variation in a Very Small Place: Social Homogeneity, Prestige Norms, and Linguistic Variation. *Language* 70(4). 631–696. doi:10.2307/416324. https://doi.org/10.2307/416324

Duan, Yuguang & Lupyan, Gary. 2023. Divergence in Word Meanings and its Consequence for Communication. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 45. https://escholarship.org/uc/item/0dp4790t.

Dummett, Michael. 1986. A Nice Derangement of Epitaphs: Some Comments on Davidson and Hacking. In LePore, Ernest (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, 459–476. Cambridge, MA: Blackwell.

Dupre, Gabe. 2021. (What) Can Deep Learning Contribute to Theoretical Linguistics? *Minds and Machines* doi:10.1007/s11023-021-09571-w. https://doi.org/10.1007/s11023-021-09571-w

Dupré, John. 1998. Normal People. *Social Research* 65(2). 221–248. https://www.jstor.org/stable/40971271.

Durkheim, Emile. 1893/1984. *The Division of Labor in Society*. New York, NY: The Free Press. https://n2t.net/ark:/13960/t8z90605p.

Durkheim, Emile. 1895/2013. The Rules of Sociological Method. In Lukes, Steven (ed.), *Emile Durkheim: The Rules of Sociological Method and Selected Texts on Sociology and its Method*, 1–120. London: Palgrave MacMillan 2nd edn.

Durkheim, Emile. 1897/2005. *Suicide: A Study in Sociology*. London and New York: Routledge. doi:10.4324/9780203994320. `https://doi.org/10.4324/9780203994320`

Durkheim, Emile. 1901/2013. Preface to the Second Edition (1901). In Lukes, Steven (ed.), *Emile Durkheim: The Rules of Sociological Method and Selected Texts on Sociology and its Method*, 6–17. London: Palgrave MacMillan 2nd edn.

Dyck, Erika & Russell, Ginny. 2020. Challenging Psychiatric Classification: Healthy Autistic Diversity and the Neurodiversity Movement. In Taylor, Steven J. & Brumby, Alice (eds.), *Healthy Minds in the Twentieth Century: In and Beyond the Asylum*, 167–187. Cham: Springer International Publishing. doi:10.1007/978-3-030-27275-3_8. `https://doi.org/10.1007/978-3-030-27275-3_8`

Eagleton, Terry. 1983/2008. *Literary Theory: An Introduction*. Minneapolis, MN: University of Minnesota Press 2nd edn. `https://n2t.net/ark:/13960/t13n3fq5w`.

Ebersbach, Georg & Sojer, Martin & Müller, Joerg & Heijmenberg, Michaela & Poewe, Werner. 2000. Sociocultural differences in gait. *Movement Disorders* 15(6). 1145–1147. doi:10.1002/1531-8257(200011)15:6<1145::AID-MDS1013>3.0.CO;2-C. `https://doi.org/10.1002/1531-8257(200011)15:6<1145::AID-MDS1013>3.0.CO;2-C`

Edelman, Gerald. 1988. *Topobiology: An Introduction To Molecular Embryology*. New York, NY: Basic Books. `https://n2t.net/ark:/13960/t23b73121`.

Edelman, Shimon & Christiansen, Morten H. 2003. How seriously should we take Minimalist syntax? *Trends in Cognitive Sciences* 7(2). 60–61. doi:10.1016/S1364-6613(02)00045-1. `https://doi.org/10.1016/S1364-6613(02)00045-1`

Enfield, Nick J. & Kockelman, Paul & Sidnell, Jack (eds.). 2014. *The Cambridge Handbook of Linguistic Anthropology* (Cambridge Handbooks in Language and Linguistics). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139342872. `https://doi.org/10.1017/CBO9781139342872`

Epstein, Robert. 1992. The Quest for the Thinking Computer. *AI Magazine* 13(2 SE - Articles). 81. doi:10.1609/aimag.v13i2.993. `https://doi.org/10.1609/aimag.v13i2.993`

Epstein, Robert & Roberts, Gary & Beber, Grace (eds.). 2009. *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Dordrecht: Springer Netherlands. doi:10.1007/978-1-4020-6710-5. `https://doi.org/10.1007/978-1-4020-6710-5`

Ereshefsky, Marc. 2022. Species. In Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Summer 2022 edn. `https://plato.stanford.edu/archives/sum2022/entries/species/`.

Everard, Andrea & Galletta, Dennis F. 2005. How Presentation Flaws Affect Perceived Site Quality, Trust, and Intention to Purchase from an Online Store. *Journal of Management Information Systems* 22(3). 56–95. doi:10.2753/MIS0742-1222220303. https://doi.org/10.2753/MIS0742-1222220303

Faruqui, Manaal & Tsvetkov, Yulia & Rastogi, Pushpendre & Dyer, Chris. 2016. Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. 30–35. Association for Computational Linguistics. doi:10.18653/v1/W16-2506. https://doi.org/10.18653/v1/W16-2506

Featherston, Sam. 2007. Data in generative grammar: The stick and the carrot. *Theoretical Linguistics* 33(3). 269–318. doi:10.1515/TL.2007.020. https://doi.org/10.1515/TL.2007.020

Federmeier, Kara D. & Kutas, Marta. 1999. A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language* 41(4). 469 – 495. doi:10.1006/jmla.1999.2660. https://doi.org/10.1006/jmla.1999.2660

Feigenbaum, Edward A. & Feldman, Julian (eds.). 1963. *Computers and Thought*. New York, NY: McGraw-Hill. https://n2t.net/ark:/13960/t1ng5m220.

Feigenbaum, Edward A. & McCorduck, Pamela. 1983. *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World*. Reading, MA: Addison-Wesley Publishing Company. https://n2t.net/ark:/13960/t7sp2tz9b.

Field, Anjalie & Blodgett, Su Lin & Waseem, Zeerak & Tsvetkov, Yulia. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1905–1925. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.149. https://aclanthology.org/2021.acl-long.149

Firth, John R. 1957/1962. A Synopsis of Linguistic Theory, 1930-1955. In Firth, John R. (ed.), *Studies in Linguistic Analysis* (Special Volume of the Philological Society), 1–32. Oxford: Basil Blackwell.

Fish, Stanley. 1980. *Is There a Text in This Class? The Authority of Interpretive Communities*. Cambridge, MA: Harvard University Press. https://n2t.net/ark:/13960/s21njh0t6bf.

Fleisig, Eve & Smith, Genevieve & Bossi, Madeline & Rustagi, Ishita & Yin, Xavier & Klein, Dan. 2024. Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination. doi:10.48550/arXiv.2406.08818. https://doi.org/10.48550/arXiv.2406.08818

Flor, Michael & Futagi, Yoko & Lopez, Melissa & Mulholland, Matthew. 2015. Patterns of misspellings in L2 and L1 English: a view from the ETS Spelling Corpus. *Bergen Language and Linguistics Studies* 6(0 SE - Articles). doi:10.15845/bells.v6i0.811. https://doi.org/10.15845/bells.v6i0.811

Fodor, Jerry A. 1975. *The Language of Thought*. New York, NY: Thomas Y. Crowell.

Fodor, Jerry A. 2008. *LOT 2: The Language of Thought Revisited*. Oxford: Clarendon Press. doi:10.1093/acprof:oso/9780199548774.001.0001. https://doi.org/10.1093/acprof:oso/9780199548774.001.0001

Forbes, Samuel H. & Aneja, Prerna & Guest, Olivia. 2022. The myth of normative development. *Infant and Child Development* e2393. doi:10.1002/icd.2393. https://doi.org/10.1002/icd.2393

Fostel, Gary. 1993. The Turing test is for the birds. *SIGART Bulletin* 4(1). 7–8. doi:10.1145/173993.173996. https://doi.org/10.1145/173993.173996

Frazier, Lyn. 1979. *On comprehending sentences: Syntactic parsing strategies*: University of Connecticut dissertation. https://digitalcommons.lib.uconn.edu/dissertations/AAI7914150/.

Frazier, Lyn & Fodor, Janet Dean. 1978. The sausage machine: A new two-stage parsing model. *Cognition* 6(4). 291–325. doi:10.1016/0010-0277(78)90002-1. https://doi.org/10.1016/0010-0277(78)90002-1

Frege, Gottlob. 1897/1979. Logic. In Hermes, Hans & Kambartel, Friedrich & Kaulbach, Friedrich (eds.), *Posthumous Writings*, 126–151. Oxford: Basil Blackwell. https://n2t.net/ark:/13960/t40s7nj3z.

Frege, Gottlob. 1914/1979. Logic in Mathematics. In Hermes, Hans & Kambartel, Friedrich & Kaulbach, Friedrich (eds.), *Posthumous Writings*, 203–250. Oxford: Basil Blackwell. https://n2t.net/ark:/13960/t40s7nj3z.

French, Robert M. 1990. Subcognition and the Limits of the Turing Test. *Mind* 99(393). 53–65. doi:10.1093/mind/XCIX.393.53. https://doi.org/10.1093/mind/XCIX.393.53

French, Robert M. 2012. Moving beyond the Turing test. *Communications of the ACM* 55(12). 74–77. doi:10.1145/2380656.2380674. https://doi.org/10.1145/2380656.2380674

Friedberg, Richard M. 1958. A Learning Machine: Part I. *IBM Journal of Research and Development* 2(1). 2–13. doi:10.1147/rd.21.0002. https://doi.org/10.1147/rd.21.0002

Friedrich, Niklas & Lauscher, Anne & Ponzetto, Simone Paolo & Glavaš, Goran. 2021. DebIE: A platform for implicit and explicit debiasing of word embedding spaces. In

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. 91–98. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.eacl-demos.11.

Froese, Tom. 2018. Searching for the Conditions of Genuine Intersubjectivity: From Agent-Based Models to Perceptual Crossing Experiments. In Newen, Albert & De Bruin, Leon & Gallagher, Shaun (eds.), *The Oxford Handbook of 4E Cognition*, 163–186. Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780198735410.013.9. https://doi.org/10.1093/oxfordhb/9780198735410.013.9

Furnas, G. W. & Landauer, T. K. & Gomez, L. M. & Dumais, S. T. 1987. The vocabulary problem in human-system communication. *Communications of the ACM* 30(11). 964–971. doi:10.1145/32206.32212. https://doi.acm.org/10.1145/32206.32212

Gallistel, Charles R. 1990. *The Organization of Learning: Development, and Conceptual Change*. Cambridge, MA: MIT Press. https://n2t.net/ark:/13960/t56f59w94.

Galton, Francis. 1869. *Hereditary Genius: An Inquiry Into Its Laws and Consequences*. London: Macmillan and Co. https://n2t.net/ark:/13960/t7wm1j30b.

Gärdenfors, Peter. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambridge, MA: MIT Press. doi:10.7551/mitpress/9629.001.0001. https://doi.org/10.7551/mitpress/9629.001.0001

Garg, Nikhil & Schiebinger, Londa & Jurafsky, Dan & Zou, James. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115(16). E3635–E3644. doi:10.1073/pnas.1720347115. https://doi.org/10.1073/pnas.1720347115

Garrod, Simon & Pickering, Martin J. 2009. Joint Action, Interactive Alignment, and Dialog. *Topics in Cognitive Science* 1(2). 292–304. doi:10.1111/j.1756-8765.2009.01020.x. https://doi.org/10.1111/j.1756-8765.2009.01020.x

Gastaldi, Juan Luis. 2021. Why Can Computers Understand Natural Language? *Philosophy & Technology* 34(1). 149–214. doi:10.1007/s13347-020-00393-9. https://doi.org/10.1007/s13347-020-00393-9

Gastaldi, Juan Luis & Pellissier, Luc. 2021. The calculus of language: explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews* 46(4). 569–590. doi:10.1080/03080188.2021.1890484. https://doi.org/10.1080/03080188.2021.1890484

Gellner, Ernest. 1964. *Thought and Change*. London: Weidenfeld and Nicolson. https://n2t.net/ark:/13960/t8kd9gr8b.

Gerz, Daniela & Vulić, Ivan & Hill, Felix & Reichart, Roi & Korhonen, Anna. 2016. SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity. In *Proceedings of* *the 2016 Conference on Empirical Methods in Natural Language Processing*. 2173–2182. Association for Computational Linguistics. doi:10.18653/v1/D16-1235. https://doi.org/10.18653/v1/D16-1235

Gibson, Edward & Fedorenko, Evelina. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14(6). 233–234. doi:10.1016/j.tics.2010.03.005. https://doi.org/10.1016/j.tics.2010.03.005

Gibson, Edward & Futrell, Richard & Piandadosi, Steven T. & Dautriche, Isabelle & Mahowald, Kyle & Bergen, Leon & Levy, Roger. 2019. How Efficiency Shapes Human Language. *Trends in Cognitive Sciences* 23(5). 389 – 407. doi:10.1016/j.tics.2019.02.003. https://doi.org/10.1016/j.tics.2019.02.003

Gibson, Edward & Piantadosi, Steven T. & Fedorenko, Evelina. 2013. Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes* 28(3). 229–240. doi:10.1080/01690965.2012.704385. https://doi.org/10.1080/01690965.2012.704385

Gilbert, Margaret. 1989. *On Social Facts*. Princeton, NJ: Princeton University Press. doi:10.2307/j.ctv10vm20z. https://doi.org/10.2307/j.ctv10vm20z

Gladwin, Thomas. 1970/2009. *East Is a Big Bird: Navigation and Logic on Puluwat Atoll*. Cambridge, MA: Harvard University Press. doi:10.2307/j.ctvjsf6g9. https://doi.org/10.2307/j.ctvjsf6g9

Goertzel, Ben & Pennachin, Cassio (eds.). 2007. *Artificial General Intelligence* (Cognitive Technologies). Berlin: Springer. doi:10.1007/978-3-540-68677-4. http://doi.org/10.1007/978-3-540-68677-4

Gopnik, Alison. 2009. *The Philosophical Baby: What Children's Minds Tell Us About Truth, Love, and the Meaning of Life*. New York, NY: Farrar, Straus and Giroux. https://n2t.net/ark:/13960/t6n057h77.

Gopnik, Myrna. 1994. Impairments of tense in a familial language disorder. *Journal of Neurolinguistics* 8(2). 109–133. doi:10.1016/0911-6044(94)90020-5. https://doi.org/10.1016/0911-6044(94)90020-5

Gould, Stephen J. & Lewontin, Richard C. 1979. The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society of London. Series B, Biological Sciences* 205(1161). 581–598. doi:10.1098/rspb.1979.0086. https://doi.org/10.1098/rspb.1979.0086

Graeber, David. 2009. *Direct Action: an Ethnography*. Oakland, CA: AK Press. https://n2t.net/ark:/13960/t9p37bq73.

Graeber, David & Wengrow, David. 2021. *The Dawn of Everything: A New History of Humanity*. New York, NY: Farrar, Straus and Giroux. https://n2t.net/ark:/13960/s21t13k7wk4.

Grant, Margaret & Sloggett, Shayne & Dillon, Brian. 2020. Processing ambiguities in attachment and pronominal reference. *Glossa: a journal of general linguistics* 5(1). 77. doi:10.5334/gjgl.852. `https://doi.org/10.5334/gjgl.852`

Greenberg, Joseph H. 1959/1974. Language and Evolution. In Cohen, Yehudi A. (ed.), *Man in Adaptation: The Cultural Present*, 33–44. New York, NY: Aldone Publishing Company. `https://n2t.net/ark:/13960/t03z5h643`.

Greenfield, Patricia M. & Savage-Rumbaugh, E. Sue. 1993. Comparing communicative competence in child and chimp: the pragmatics of repetition. *Journal of Child Language* 20(1). 1–26. doi:10.1017/S0305000900009090. `https://doi.org/10.1017/S0305000900009090`

Grice, H. Paul. 1957. Meaning. *The Philosophical Review* 66(3). 377–388. doi:10.2307/2182440. `https:/doi.org/10.2307/2182440`

Grice, H. Paul. 1968. Utterer's Meaning, Sentence-Meaning, and Word-Meaning. *Foundations of Language* 4(3). 225–242. `https://www.jstor.org/stable/25000329`.

Grice, H. Paul. 1975. Logic and Conversation. In *Speech Acts*, 41–58. Leiden: Brill. doi:10.1163/9789004368811_003. `https://doi.org//10.1163/9789004368811_003`

Grice, H. Paul. 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press. `https://n2t.net/ark:/13960/t0dv5nd9m`.

Gutting, Gary & Oksala, Johanna. 2022. Michel Foucault. In Zalta, Edward N. & Nodelman, Uri (eds.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Fall 2022 edn. `https://plato.stanford.edu/archives/fall2022/entries/foucault/`.

Hacking, Ian. 1990. *The Taming of Chance*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511819766. `https://doi.org/10.1017/CBO9780511819766`

Hald, Anders. 1986. Galileo's Statistical Analysis of Astronomical Observations. *International Statistical Review / Revue Internationale de Statistique* 54(2). 211–220. doi:10.2307/1403145. `https://doi.org/10.2307/1403145`

Halmos, Paul R. 1958/1987. *Finite-Dimensional Vector Spaces* (Undergraduate Texts in Mathematics). New York, NY: Springer-Verlag. doi:10.1007/978-1-4612-6387-6. `https://doi.org/10.1007/978-1-4612-6387-6`

Halmos, Paul R. 1974. *Naive Set Theory* (Undergraduate Texts in Mathematics). New York, NY: Springer-Verlag. doi:10.1007/978-1-4757-1645-0. `https://doi.org/10.1007/978-1-4757-1645-0`

Harnad, Stevan. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42(1). 335–346. doi:10.1016/0167-2789(90)90087-6. `https://doi.org/10.1016/0167-2789(90)90087-6`

Harnad, Stevan. 1992. The Turing Test is Not a Trick: Turing Indistinguishability is a Scientific Criterion. *SIGART Bulletin* 3(4). 9–10. doi:10.1145/141420.141422. `https://doi.org/10.1145/141420.141422`

Harris, Zellig S. 1954. Distributional Structure. *Word* 10(2-3). 146–162. doi:10.1080/00437956.1954.11659520. `https://doi.org/10.1080/00437956.1954.11659520`

Hassan, Saad & Huenerfauth, Matt & Alm, Cecilia Ovesdotter. 2021. Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems through an Intersectional Lens. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 3116–3123. Punta Cana, Dominican Republic: Association for Computational Linguistics. doi:10.18653/v1/2021.findings-emnlp.267. `https://doi.org/10.18653/v1/2021.findings-emnlp.267`

Hayes, Patrick & Ford, Kenneth. 1995. Turing Test Considered Harmful. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1* (IJCAI'95). 972–977. San Francisco, CA: Morgan Kaufmann Publishers Inc. doi:10.5555/1625855.1625981. `https://doi.org/10.5555/1625855.1625981`

Henley, Nancy M. 1969. A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior* 8(2). 176–184. doi:10.1016/S0022-5371(69)80058-7. `https://doi.org/10.1016/S0022-5371(69)80058-7`

Henrich, Joseph & Heine, Steven J. & Norenzayan, Ara. 2010a. Most people are not WEIRD. *Nature* 466(7302). 29. doi:10.1038/466029a. `https://doi.org/10.1038/466029a`

Henrich, Joseph & Heine, Steven J. & Norenzayan, Ara. 2010b. The weirdest people in the world? *Behavioral and Brain Sciences* 33(2-3). 61–83. doi:10.1017/S0140525X0999152X. `https://doi.org/10.1017/S0140525X0999152X`

Herbelot, Aurélie & von Redecker, Eva & Müller, Johanna. 2012. Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. 45–54. Avignon, France: Association for Computational Linguistics. `https://aclanthology.org/W12-1008`.

Hewes, Gordon W. 1961. Food Transport and the Origin of Hominid Bipedalism. *American Anthropologist* 63(4). 687–710. doi:10.1525/aa.1961.63.4.02a00020. `https://doi.org/10.1525/aa.1961.63.4.02a00020`

Hill, Benjamin. 2021. Locke on the origins, nature, and functions of language. In Gordon-Roth, Jessica & Weinberg, Shelley (eds.), *The Lockean Mind* (Routledge Philosophical Minds), 167–180. New York and London: Routledge. doi:10.4324/9781315099675. `https://doi.org/10.4324/9781315099675`

Hill, Felix & Reichart, Roi & Korhonen, Anna. 2015. Simlex-999: Evaluating Semantic Models with Genuine Similarity Estimation. *Computational Linguistics* 41(4). 665–695. doi:10.1162/COLI_a_00237. https://doi.org/10.1162/COLI_a_00237

Hobbes, Thomas. 1651/2010. *Leviathan: Or the Matter, Forme and Power of a Commonwealth, Ecclesiasticall and Civill* (Rethinking the Western Tradition). New Haven, CT: Yale University Press.

Hodges, Andrew. 1983/2014. *Alan Turing: The Enigma.* Princeton, NJ: Princeton University Press revised edn. doi:10.2307/j.ctvc77913. https://doi.org/10.2307/j.ctvc77913

Hofstadter, Douglas R. 1985. *Metamagical themas: Questing for the Essence of Mind and Pattern.* New York, NY: Basic Books. https://n2t.net/ark:/13960/t4dp1zv4m.

Holdcroft, David. 1991. *Saussure: Signs, System and Arbitrariness.* Cambridge, UK: Cambridge University Press. https://n2t.net/ark:/13960/s2hjr3bnd3c.

Horn, John L. 1979. Trends in the measurement of intelligence. *Intelligence* 3(3). 229–239. doi:10.1016/0160-2896(79)90019-9. https://doi.org/10.1016/0160-2896(79)90019-9

Huang, Yan. 2012. *The Oxford Dictionary of Pragmatics.* Oxford: Oxford University Press.

Hull, David L. 1986. On Human Nature. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1986. 3–13. doi:10.1086/psaprocbienmeetp.1986.2.192787. https://doi.org/10.1086/psaprocbienmeetp.1986.2.192787

Humboldt, Wilhelm von. 1836/1988. *On Language: The Diversity of Human Language-Structure and its Influence on the Mental Development of Mankind.* Cambridge, UK: Cambridge University Press.

Humpherys, Jeffrey & Jarvis, Tyler J. & Evans, Emily J. 2017. *Foundations of Applied Mathematics, Volume 1: Mathematical Analysis.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Hutchens, Jason L. 2009. Conversation Simulation and Sensible Surprises. In Epstein, Robert & Roberts, Gary & Beber, Grace (eds.), *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, 325–342. Dordrecht: Springer Netherlands. doi:10.1007/978-1-4020-6710-5_20. https://doi.org/10.1007/978-1-4020-6710-5_20

Hutchinson, J. Wesley & Lockhead, Gregory R. 1977. Similarity as distance: A structural principle for semantic memory. *Journal of Experimental Psychology: Human Learning and Memory* 660–678. doi:10.1037/0278-7393.3.6.660. https://doi.org/10.1037/0278-7393.3.6.660

Hutchinson, John R. & Gatesy, Stephen M. 2001. Bipedalism. In *Encyclopedia of Life Sciences*, New York, NY: John Wiley & Sons. doi:10.1038/npg.els.0001869. https://doi.org/10.1038/npg.els.0001869

Hutton, Christopher. 2009. *Language, Meaning and the Law*. Edinburgh: Edinburgh University Press. doi:10.1515/9780748633524. https://doi.org/10.1515/9780748633524

Huxley, Thomas Henry. 1863/2009. *Evidence as to Man's Place in Nature*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511703539. https://doi.org/10.1017/CBO9780511703539

Ingold, Tim. 2004. Culture on the Ground: The World Perceived Through the Feet. *Journal of Material Culture* 9(3). 315–340. doi:10.1177/1359183504046896. https://doi.org/10.1177/1359183504046896

Ingold, Tim & Vergunst, Jo Lee (eds.). 2008. *Ways of Walking: Ethnography and Practice on Foot* (Anthropological Studies of Creativity and Perception). London and New York: Routledge. doi:10.4324/9781315234250. https://doi.org/10.4324/9781315234250

Itkonen, Esa. 1978. *Grammatical Theory and Metascience: A critical investigation into the methodological and philosophical foundations of 'autonomous' linguistics* (Current Issues in Linguistic Theory 5). Amsterdam: John Benjamins Publishing Company. doi:10.1075/cilt.5. https://doi.org/10.1075/cilt.5

Itkonen, Esa. 2008. The central role of normativity in language and linguistics. In Zlatev, Jordan & Racine, Timothy P. & Sinha, Chris & Itkonen, Esa (eds.), *The Shared Mind: Perspectives on intersubjectivity*, 279–305. Amsterdam: John Benjamins Publishing Company. doi:10.1075/celcr.12.16itk. https://doi.org/10.1075/celcr.12.16itk

Jackman, Henry. 2020. Meaning Holism. In Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University winter 2020 edn. https://plato.stanford.edu/archives/win2020/entries/meaning-holism/.

Jelinek, Frederick. 2005. Some of My Best Friends Are Linguists. *Language Resources and Evaluation* 39(1). 25–34. doi:10.1007/s10579-005-2693-4. https://doi.org/10.1007/s10579-005-2693-4

Jeong, Allan & Li, Haiying & Pan, Andy Jiaren. 2017. A sequential analysis of responses in online debates to postings of students exhibiting high versus low grammar and spelling errors. *Educational Technology Research and Development* 65(5). 1175–1194. doi:10.1007/s11423-016-9501-2. https://doi.org/10.1007/s11423-016-9501-2

Johnson, Samuel. 1747/2021. The Plan of A Dictionary of the English Language. In DeMaria, Robert & Fix, Stephen & Weinbrot, Howard D. (eds.), *Samuel Johnson: Selected Works*, 378–396. New Haven, CT: Yale University Press. doi:10.12987/9780300258004-037. https://doi.org/10.12987/9780300258004-037

Johnstone, Barbara. 2000. The Individual Voice in Language. *Annual Review of Anthropology* 29(1). 405–424. doi:10.1146/annurev.anthro.29.1.405. https://doi.org/10.1146/annurev.anthro.29.1.405

Jones, Cameron R. & Bergen, Benjamin K. 2024. People cannot distinguish GPT-4 from a human in a Turing test. doi:10.48550/arXiv.2405.08007. https://doi.org/10.48550/arXiv.2405.08007

Jurafsky, Daniel & Martin, James H. 2022. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Third Edition draft). https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf.

Kachru, Braj B. & Kahane, Henry (eds.). 1995. *Cultures, Ideologies, and the Dictionary: Studies in Honor of Ladislav Zgusta*. Berlin and Boston: Max Niemeyer Verlag. doi: doi:10.1515/9783110957075. https://doi.org/10.1515/9783110957075

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux. https://n2t.net/ark:/13960/t0200hg3b.

Kaneko, Masahiro & Bollegala, Danushka. 2021a. Debiasing Pre-trained Contextualised Embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1256–1266. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.eacl-main.107.

Kaneko, Masahiro & Bollegala, Danushka. 2021b. Dictionary-based debiasing of pre-trained word embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 212–223. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.eacl-main.16.

Kegl, Judy & Senghas, Ann & Coppola, Marie. 1999. Creation through contact: Sign language emergence and sign language change in Nicaragua. In DeGraff, Michel (ed.), *Language Creation and Language Change: Creolization, Diachrony, and Development*, Cambridge, MA: MIT Press.

Kemmerer, David. 2014. *Cognitive Neuroscience of Language*. New York and London: Psychology Press. doi:10.4324/9781315764061. https://doi.org/10.4324/9781315764061

Keucheyan, Razmig. 2010. Durkheim, Wittgenstein and the Norms of Thought. *Diogenes* 57(4). 63–72. doi:10.1177/0392192111435749. https://doi.org/10.1177/0392192111435749

Keysar, Boaz & Barr, Dale J. & Horton, William S. 1998. The Egocentric Basis of Language Use: Insights From a Processing Approach. *Current Directions in Psychological*

Science 7(2). 46–49. doi:10.1111/1467-8721.ep13175613. https://doi.org/10.1111/1467-8721.ep13175613

Keysar, Boaz & Henly, Anne S. 2002. Speakers' Overestimation of Their Effectiveness. *Psychological Science* 13(3). 207–212. doi:10.1111/1467-9280.00439. https://doi.org/10.1111/1467-9280.00439

Keysar, Boaz & Lin, Shuhong & Barr, Dale J. 2003. Limits on theory of mind use in adults. *Cognition* 89(1). 25–41. doi:10.1016/S0010-0277(03)00064-7. https://doi.org/10.1016/S0010-0277(03)00064-7

King, Stephen. 2014. *On Writing: A Memoir Of The Craft*. New York, NY: Scribner. https://n2t.net/ark:/13960/t51h3170b.

Korta, Kepa & Perry, John. 2020. Pragmatics. In Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University spring 2020 edn. https://plato.stanford.edu/archives/spr2020/entries/pragmatics/.

Kraljic, Tanya & Brennan, Susan E. 2005. Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology* 50(2). 194–231. doi:10.1016/j.cogpsych.2004.08.002. https://doi.org/10.1016/j.cogpsych.2004.08.002

Kreiner, David S. & Schnakenberg, Summer D. & Green, Angela G. & Costello, Michael J. & McClin, Anis F. 2002. Effects of Spelling Errors on the Perception of Writers. *The Journal of General Psychology* 129(1). 5–17. doi:10.1080/00221300209602029. https://doi.org/10.1080/00221300209602029

Kremer, Gerhard & Erk, Katrin & Padó, Sebastian & Thater, Stefan. 2014. What Substitutes Tell Us - Analysis of an "All-Words" Lexical Substitution Corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 540–549. Gothenburg, Sweden: Association for Computational Linguistics. doi:10.3115/v1/E14-1057. https://doi.org/10.3115/v1/E14-1057

Kruger, Justin & Epley, Nicholas & Parker, Jason & Ng, Zhi-Wen. 2005. Egocentrism Over E-Mail: Can We Communicate as Well as We Think? *Journal of Personality and Social Psychology* 89(6). 925–936. doi:10.1037/0022-3514.89.6.925. https://doi.org/10.1037/0022-3514.89.6.925

Kummer, Hans & Goodall, Jane. 1985. Conditions of Innovative Behaviour in Primates. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 308(1135). 203–214. doi:10.1098/rstb.1985.0020. https://doi.org/10.1098/rstb.1985.0020

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press.

Labov, William. 1973. The boundaries of words and their meanings. In Bailey, Charles-James N. & Shuy, Roger W. (eds.), *New Ways of Analyzing Variation in English*, 340–373. Washington, DC: Georgetown University Press. `https://n2t.net/ark:/13960/t58d7f235`.

Labov, William. 1994. *Principles of Linguistic Change, Volume 1: Internal Factors*. Oxford: Wiley-Blackwell.

Labov, William. 2001. *Principles of Linguistic Change, Volume 2: Social Factors*. Oxford: Wiley-Blackwell.

Labov, William. 2010. *Principles of Linguistic Change, Volume 3: Cognitive and Cultural Factors*. Oxford: Wiley-Blackwell. doi:10.1002/9781444327496. `https://doi.org/10.1002/9781444327496`

Laher, Sumaya & Cockcroft, Kate. 2014. Psychological assessment in post-apartheid South Africa: the way forward. *South African Journal of Psychology* 44(3). 303–314. doi:10.1177/0081246314533634. `https://doi.org/10.1177/0081246314533634`

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar. Volume I: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Lau, Becky Ka Ying & Geipel, Janet & Wu, Yanting & Keysar, Boaz. 2022. The extreme illusion of understanding. *Journal of Experimental Psychology: General* 151. 2957–2962. doi:10.1037/xge0001213. `https://doi.org/10.1037/xge0001213`

Lauscher, Anne & Lueken, Tobias & Glavaš, Goran. 2021. Sustainable Modular Debiasing of Language Models. In Moens, Marie-Francine & Huang, Xuanjing & Specia, Lucia & Yih, Scott Wen-tau (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4782–4797. Punta Cana, Dominican Republic: Association for Computational Linguistics. doi:10.18653/v1/2021.findings-emnlp.411. `https://doi.org/10.18653/v1/2021.findings-emnlp.411`

Le Page, Robert B. & Tabouret-Keller, Andrée. 1985. *Acts of Identity: Creole-Based Approaches to Language and Ethnicity*. Cambridge, UK: Cambridge University Press.

Le Page, Robert B. & Tabouret-Keller, Andrée. 2010. Acts of Identity. In Griffiths, Patrick & Merrison, Andrew John & Bloomer, Aileen (eds.), *Language in Use: A Reader*, 183–189. London and New York: Routledge. doi:10.4324/9781003060994. `https://doi.org/10.4324/9781003060994`

Lenci, Alessandro. 2018. Distributional Models of Word Meaning. *Annual review of Linguistics* 4. 151–171. doi:10.1146/annurev-linguistics-030514-125254. `https://doi.org/10.1146/annurev-linguistics-030514-125254`

Lenci, Alessandro & Sahlgren, Magnus & Jeuniaux, Patrick & Cuba Gyllensten, Amaru & Miliani, Martina. 2022. A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language Resources and Evaluation* doi:10.1007/s10579-021-09575-z. https://doi.org/10.1007/s10579-021-09575-z

Levinson, Stephen C. 1983. *Pragmatics* (Cambridge Textbooks in Linguistics). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511813313. https://doi.org/10.1017/CBO9780511813313

Levshina, Natalia. 2021. Cross-Linguistic Trade-Offs and Causal Relationships Between Cues to Grammatical Subject and Object, and the Problem of Efficiency-Related Explanations. *Frontiers in Psychology* 12. doi:10.3389/fpsyg.2021.648200. https://doi.org/10.3389/fpsyg.2021.648200

Levshina, Natalia. 2022. *Communicative Efficiency: Language Structure and Use*. Cambridge, UK: Cambridge University Press. doi:10.1017/9781108887809. https://doi.org/10.1017/9781108887809

Lewis, David. 1969/2002. *Convention: a Philosophical Study*. Oxford: Blackwell. doi:10.1002/9780470693711. https://doi.org/10.1002/9780470693711

Lewis, David. 1975/1983. Languages and Language. In *Philosophical Papers Volume I*, 163–188. Oxford: Oxford University Press. doi:10.1093/0195032047.003.0011. https://doi.org/10.1093/0195032047.003.0011

Lewontin, Richard C. 1998. The evolution of cognition: Questions we will never answer. In *Methods, models, and conceptual issues: An invitation to cognitive science, vol. 4.* (An invitation to cognitive science.), 106–132. Cambridge, MA: MIT Press. doi:10.7551/mitpress/3967.003.0006. https://doi.org/10.7551/mitpress/3967.003.0006

Lieberman, Daniel E. & Venkadesan, Madhusudhan & Werbel, William A. & Daoud, Adam I. & D'Andrea, Susan & Davis, Irene S. & Mang'Eni, Robert Ojiambo & Pitsiladis, Yannis. 2010. Foot strike patterns and collision forces in habitually barefoot versus shod runners. *Nature* 463(7280). 531–535. doi:10.1038/nature08723. https://doi.org/10.1038/nature08723

Linell, Per. 2017. Intersubjectivity in Dialogue. In Weigand, Edda (ed.), *The Routlege Handbook of Language and Dialogue*, 109–126. New York, NY: Routledge. doi:10.4324/9781315750583. https://doi.org/10.4324/9781315750583

Liu, Ziming. 2004. Perceptions of credibility of scholarly information on the web. *Information Processing & Management* 40(6). 1027–1038. doi:10.1016/S0306-4573(03)00064-5. https://doi.org/10.1016/S0306-4573(03)00064-5

Locke, John. 1690/2004. *An Essay Concerning Human Understanding*. London: Penguin Books. https://n2t.net/ark:/13960/s2xn75wgcvn.

Loftus, Elizabeth F. 1979. *Eyewitness Testimony*. Cambridge, MA: Harvard University Press. https://n2t.net/ark:/13960/t9p35gv3v.

Longrich, Nicholas R. & Vinther, Jakob & Meng, Qingjin & Li, Quangguo & Russell, Anthony P. 2012. Primitive Wing Feather Arrangement in *Archaeopteryx lithographica* and *Anchiornis huxleyi*. *Current Biology* 22(23). 2262–2267. doi:10.1016/j.cub.2012.09.052. https://doi.org/10.1016/j.cub.2012.09.052

Ludlow, Peter. 2014. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780198712053.001.0001. https://doi.org/10.1093/acprof:oso/9780198712053.001.0001

Lukes, Steven. 1982/2013. Introduction to this Edition. In Lukes, Steven (ed.), *Emile Durkheim: The Rules of Sociological Method and Selected Texts on Sociology and its Method*, xi–xxxv. London: Palgrave MacMillan 2nd edn.

Lunsford, Andrea A. & Lunsford, Karen J. 2008. "Mistakes Are a Fact of Life": A National Comparative Study. *College Composition and Communication* 59(4). 781–806. https://www.jstor.org/stable/20457033.

Lupyan, Gary & Piantadosi, Steven T. & Marti, Louis & Kidd, Celeste & Liu, Ella & Van Paridon, Jeroen & Bi, Yanchao. 2022. How Much Do We Agree on What Words Mean? In Ravignani, A. & Asano, R. & Valente, D. & Ferretti, F. & Hartmann, S. & Hayashi, M. & Jadoul, Y. & Martins, M. & Oseki, Y. & Rodrigues, E. D. & Vasileva, O. & Wacewicz, S. (eds.), *Proceedings of the Joint Conference on Language Evolution (JCoLE)*. doi:10.17617/2.3398549. https://evolang.org/jcole2022/proceedings/paper.html?nr=140

Mac Lane, Saunders. 1996. Structure in mathematics. *Philosophia Mathematica* 4(2). 174–183. doi:10.1093/philmat/4.2.174. https://doi.org/10.1093/philmat/4.2.174

MacKenzie, Donald A. 1981. *Statistics in Britain, 1865–1930: The Social Construction of Scientific Knowledge*. Edinburgh: Edinburgh University Press.

Malt, Barbara C. 2020. Words, thoughts, and brains. *Cognitive Neuropsychology* 37(5-6). 241–253. doi:10.1080/02643294.2019.1599335. https://doi.org/10.1080/02643294.2019.1599335

Malt, Barbara C. & Li, Ping & Pavlenko, Aneta & Zhu, Huichun & Ameel, Eef. 2015. Bidirectional lexical interaction in late immersed Mandarin-English bilinguals. *Journal of Memory and Language* 82. 86–104. doi:10.1016/j.jml.2015.03.001. https://doi.org/10.1016/j.jml.2015.03.001

Maniglier, Patrice. 2011. Processing Cultures: "Structuralism" in the History of Artificial Intelligence. In *The Search for a Theory of Cognition: Early Mechanisms and New Ideas*, 145–172. Leiden: Brill | Rodopi. doi:10.1163/9789401207157_006. https://doi.org/10.1163/9789401207157_006

Marconi, Diego. 1997. *Lexical Competence* (Language, Speech, and Communication). Cambridge, MA: MIT Press.

Marcus, Gary F. 2001. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press. doi:10.7551/mitpress/1187.001.0001. https://doi.org/10.7551/mitpress/1187.001.0001

Martí, Louis & Piantadosi, Steven T. & Kidd, Celeste. 2019. Same Words, Same Context, Different Meanings: People are unaware their own concepts are not always shared. In Goel, Ashok K. & Seifert, Colleen M. & Freksa, Christian (eds.), *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019*. 2296–2302. Cognitive Science Society. https://mindmodeling.org/cogsci2019/papers/0398/index.html.

Martí, Louis & Wu, Shengyi & Piantadosi, Steven T. & Kidd, Celeste. 2023. Latent Diversity in Human Concepts. *Open Mind* 7. 79–92. doi:10.1162/opmi_a_00072. https://doi.org/10.1162/opmi_a_00072

Matthews, Gareth & Mullin, Amy. 2023. The Philosophy of Childhood. In Zalta, Edward N. & Nodelman, Uri (eds.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Fall 2023 edn. https://plato.stanford.edu/archives/fall2023/entries/childhood/.

Mauss, Marcel. 1934/2007. Techniques of the Body. In Lock, Margaret & Farquhar, Judith (eds.), *Beyond the Body Proper: Reading the Anthropology of Material Life*, 50–68. Durham, NC: Duke University Press. https://n2t.net/ark:/13960/t6p06wv8r.

Mayer, André. 1937. La vie. In Febvre, Lucien (ed.), *Encyclopédie française*, vol. 4, Paris: Société de gestion de l'encyclopédie française. https://gallica.bnf.fr/ark:/12148/bpt6k39438j.

McCarthy, Diana & Navigli, Roberto. 2009. The English lexical substitution task. *Language Resources and Evaluation* 43(2). 139–159. doi:10.1007/s10579-009-9084-1. https://doi.org/10.1007/s10579-009-9084-1

McCarthy, John. 2007. From here to human-level AI. *Artificial Intelligence* 171(18). 1174–1182. doi:10.1016/j.artint.2007.10.009. https://doi.org/10.1016/j.artint.2007.10.009

McCarthy, John & Minsky, Marvin L. & Rochester, Nathaniel & Shannon, Claude E. 1955/2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine* 27(4 SE - Articles). 12. doi:10.1609/aimag.v27i4.1904. https://doi.org/10.1609/aimag.v27i4.1904

McConnell, David & Phelan, Shanon. 2022. The devolution of eugenic practices: Sexual and reproductive health and oppression of people with intellectual disability. *Social Science & Medicine* 298. 114877. doi:10.1016/j.socscimed.2022.114877. https://doi.org/10.1016/j.socscimed.2022.114877

McQuillan, Dan. 2018. Data Science as Machinic Neoplatonism. *Philosophy & Technology* 31(2). 253–272. doi:10.1007/s13347-017-0273-3. https://doi.org/10.1007/s13347-017-0273-3

Meillet, Antoine. 1921/1982. *Linguistique historique et linguistique générale*. Paris: Champion. https://gallica.bnf.fr/ark:/12148/bpt6k21349s.

Meisel, Perry & Saussy, Haun. 2011. Notes. In Meisel, Perry & Saussy, Haun (eds.), *Course in General Linguistics: Ferdinand de Saussure*, 235–238. New York, NY: Columbia University Press.

Mikolajczyk, Tadeusz & Mikołajewska, Emilia & Al-Shuka, Hayder F. N. & Malinowski, Tomasz & Kłodowski, Adam & Pimenov, Danil Yurievich & Paczkowski, Tomasz & Hu, Fuwen & Giasin, Khaled & Mikołajewski, Dariusz & Macko, Marek. 2022. Recent Advances in Bipedal Walking Robots: Review of Gait, Drive, Sensors and Control Systems. *Sensors* 22(12). 4440. doi:10.3390/s22124440. https://doi.org/10.3390/s22124440

Mikolov, Tomas & Chen, Kai & Corrado, Greg & Dean, Jeffrey. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations*. https://openreview.net/forum?id=idpCdOWtqXd60.

Miller, George A. 1951. *Language and Communication*. New York, NY: McGraw-Hill. https://n2t.net/ark:/13960/t76t1hm7w.

Miller, George A. & Charles, Walter G. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1). 1–28. doi:10.1080/01690969108406936. https://doi.org/10.1080/01690969108406936

Miller, George A. & Isard, Stephen. 1964. Free Recall of Self-Embedded English Sentences. *Information and Control* 7(3). 292–303. doi:10.1016/S0019-9958(64)90310-9. https://doi.org/10.1016/S0019-9958(64)90310-9

Millikan, Ruth G. 2013. Natural information, intentional signs and animal communication. In Stegmann, Ulrich E. (ed.), *Animal Communication Theory: Information and Influence*, 133–148. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139003551.008. https://doi.org/10.1017/CBO9781139003551.008

Millikan, Ruth Garrett. 1984/2001. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press. doi:10.7551/mitpress/4124.001.0001. https://doi.org/10.7551/mitpress/4124.001.0001

Millikan, Ruth Garrett. 2000. *On Clear and Confused Ideas: An Essay about Substance Concepts* (Cambridge Studies in Philosophy). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511613296. https://doi.org/10.1017/CBO9780511613296

Millikan, Ruth Garrett. 2003. In Defense of Public Language. In Antony, Louise M. & Hornstein, Norbert (eds.), *Chomsky and His Critics*, 215–237. Oxford: Blackwell. doi:10.1002/9780470690024.ch9. `https://doi.org/10.1002/9780470690024.ch9`

Minsky, Marvin L. & Singh, Push & Sloman, Aaron. 2004. The St. Thomas Common Sense Symposium: Designing Architectures for Human-Level Intelligence. *AI Magazine* 25(2 SE - Articles). 113. doi:10.1609/aimag.v25i2.1764. `https://doi.org/10.1609/aimag.v25i2.1764`

Mitchell, Jeff & Bowers, Jeffrey. 2020. Priorless Recurrent Networks Learn Curiously. In Scott, Donia & Bel, Nuria & Zong, Chengqing (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*. 5147–5158. Barcelona, Spain (Online): International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.451. `https://doi.org/10.18653/v1/2020.coling-main.451`

Moro, Andrea. 2016. *Impossible Languages*. Cambridge, MA: MIT Press. doi:10.7551/mitpress/9780262034890.001.0001. `https://doi.org/10.7551/mitpress/9780262034890.001.0001`

Moro, Andrea & Greco, Matteo & Cappa, Stefano F. 2023. Large languages, impossible languages and human brains. *Cortex* 167. 82–85. doi:10.1016/j.cortex.2023.07.003. `https://doi.org/10.1016/j.cortex.2023.07.003`

Morris, Meredith Ringel & Counts, Scott & Roseway, Asta & Hoff, Aaron & Schwarz, Julia. 2012. Tweeting is believing? understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (CSCW '12). 441–450. New York, NY: Association for Computing Machinery. doi:10.1145/2145204.2145274. `https://doi.org/10.1145/2145204.2145274`

Mosier, Charles I. 1941. A Psychometric Study of Meaning. *The Journal of Social Psychology* 13(1). 123–140. doi:10.1080/00224545.1941.9714065. `https://doi.org/10.1080/00224545.1941.9714065`

Mozafari, Ahva & El-Alayli, Amani & Kunemund, Adrian & Fry, Trevor. 2019. Impressions of businesses with language errors in print advertising: Do spelling and grammar influence the inclination to use a business? *Current Psychology* 38(6). 1721–1727. doi:10.1007/s12144-017-9735-0. `https://doi.org/10.1007/s12144-017-9735-0`

Mühlenbernd, Roland. 2021. Evolutionary stability of ambiguity in context signaling games. *Synthese* 198(12). 11725–11753. doi:10.1007/s11229-020-02826-6. `https://doi.org/10.1007/s11229-020-02826-6`

Muller, Max. 1870. Darwinism tested by the Science of Language. Translated from the German of Professor August Schleicher. *Nature* 1(10). 256–259. doi:10.1038/001256a0. `https://doi.org/10.1038/001256a0`

Murphy, Gregory L. 2002. *The Big Book of Concepts*. Cambridge, MA: MIT Press. doi:10.7551/mitpress/1602.001.0001. https://doi.org/10.7551/mitpress/1602.001.0001

Musso, Mariacristina & Moro, Andrea & Glauche, Volkmar & Rijntjes, Michel & Reichenbach, Jürgen & Büchel, Christian & Weiller, Cornelius. 2003. Broca's area and the language instinct. *Nature Neuroscience* 6(7). 774–781. doi:10.1038/nn1077. https://doi.org/10.1038/nn1077

Napier, John R. 1967. The antiquity of human walking. *Scientific American* 216(4). 56–66. doi:10.1038/scientificamerican0467-56. https://doi.org/10.1038/scientificamerican0467-56

Napier, John R. 2019. Bipedalism. In Henke, Winfried & Tattersall, Ian (eds.), *Encyclopedia Britannica*, Chicago, IL: Encyclopædia Britannica. https://www.britannica.com/science/bipedalism.

Navigli, Roberto & Conia, Simone & Ross, Björn. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality* 15(2). doi:10.1145/3597307. https://doi.org/10.1145/3597307

Nefdt, Ryan M. 2020. A Puzzle concerning Compositionality in Machines. *Minds and Machines* 30(1). 47–75. doi:10.1007/s11023-020-09519-6. https://doi.org/10.1007/s11023-020-09519-6

Neisser, Ulric (ed.). 1976. *Cognition and Reality: Principles and Implications of Cognitive Psychology*. San Francisco, CA: W.H. Freeman. https://n2t.net/ark:/13960/t9t15hg3w.

Neisser, Ulric (ed.). 1982. *Memory Observed: Remembering in Natural Contexts*. San Francisco, CA: W.H. Freeman. https://n2t.net/ark:/13960/t5x64m96h.

Newell, Allen & Simon, Herbert A. 1963. GPS, a program that simulates human thought. In Feigenbaum, Edward A. & Feldman, Julian (eds.), *Computers and Thought*, 279–293. New York, NY: McGraw-Hill. https://n2t.net/ark:/13960/t1ng5m220.

Nickerson, Raymond S. 1999. How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin* 125(6). 737–759. doi:10.1037/0033-2909.125.6.737. https://doi.org/10.1037/0033-2909.125.6.737

Nilsson, Nils J. 2005. Human-Level Artificial Intelligence? Be Serious! *AI Magazine* 26(4 SE - Articles). 68. doi:10.1609/aimag.v26i4.1850. https://doi.org/10.1609/aimag.v26i4.1850

Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511819346. https://doi.org/10.1017/CBO9780511819346

Norvig, Peter. 2012. Colorless Green Ideas Learn Furiously: Chomsky and the Two Cultures of Statistical Learning. *Significance* 9(4). 30–33. doi:10.1111/j.1740-9713.2012.00590.x. `https://doi.org/10.1111/j.1740-9713.2012.00590.x`

Ochs, Elinor & Schieffelin, Bambi. 1995/2006. The impact of language socialization on grammatical development. In Jourdan, Christine & Tuite, Kevin (eds.), *Language, Culture, and Society: Key Topics in Linguistic Anthropology* (Studies in the Social and Cultural Foundations of Language), 168–189. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511616792.009. `https://doi.org/10.1017/CBO9780511616792.009`

O'Connor, Cailin. 2015. Ambiguity Is Kinda Good Sometimes. *Philosophy of Science* 82(1). 110–121. doi:10.1086/679180. `https://doi.org/10.1086/679180`

OpenAI. 2023a. ChatGPT. Accessed August 23 2023. `https://chat.openai.com/`.

OpenAI. 2023b. GPT-4 Technical Report. doi:10.48550/arXiv.2303.08774. `https://doi.org/10.48550/arXiv.2303.08774`

Origgi, Gloria & Sperber, Dan. 2000. Evolution, communication and the proper function of language. In Chamberlain, Andrew & Carruthers, Peter (eds.), *Evolution and the Human Mind: Modularity, Language and Meta-Cognition*, 140–169. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511611926.008. `https://doi.org/10.1017/CBO9780511611926.008`

Osgood, Charles E. 1952. The Nature and Measurement of Meaning. *Psychological Bulletin* 49(3). 197–237. doi:10.1037/h0055737. `https://doi.org/10.1037/h0055737`

Osgood, Charles E. & Suci, George J. & Tannenbaum, Percy H. 1957. *The Measurement of Meaning*. Urbana, IL: University of Illinois Press. `https://n2t.net/ark:/13960/t14n0724k`.

Ouyang, Long & Wu, Jeffrey & Jiang, Xu & Almeida, Diogo & Wainwright, Carroll & Mishkin, Pamela & Zhang, Chong & Agarwal, Sandhini & Slama, Katarina & Ray, Alex & Schulman, John & Hilton, Jacob & Kelton, Fraser & Miller, Luke & Simens, Maddie & Askell, Amanda & Welinder, Peter & Christiano, Paul F. & Leike, Jan & Lowe, Ryan. 2022. Training language models to follow instructions with human feedback. In Koyejo, S. & Mohamed, S. & Agarwal, A. & Belgrave, D. & Cho, K. & Oh, A. (eds.), *Advances in Neural Information Processing Systems*, vol. 35. 27730–27744. Curran Associates, Inc. `https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf`.

Page, Scott E. 2007. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton, NJ: Princeton University Press. doi:10.2307/j.ctt7sp9c. `https://doi.org/10.2307/j.ctt7sp9c`

Pan, Steven C. & Rickard, Timothy C. & Bjork, Robert A. 2021. Does Spelling Still Matter—and If So, How Should It Be Taught? Perspectives from Contemporary and Historical Research. *Educational Psychology Review* 33(4). 1523–1552. doi:10.1007/s10648-021-09611-y. https://doi.org/10.1007/s10648-021-09611-y

Pantel, Patrick. 2005. Inducing Ontological Co-occurrence Vectors. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. 125–132. Ann Arbor, MI: Association for Computational Linguistics. doi:10.3115/1219840.1219856. https://doi.org/10.3115/1219840.1219856

Paroubek, Patrick & Chaudiron, Stéphane & Hirschman, Lynette. 2007. Principles of Evaluation in Natural Language Processing. *Traitement Automatique des Langues* 48(1). 7–31. https://hal.archives-ouvertes.fr/hal-00502700.

Pelletier, Francis Jeffry. 2017. Compositionality and Concepts—A Perspective from Formal Semantics and Philosophy of Language. In Hampton, James A. & Winter, Yoad (eds.), *Compositionality and Concepts in Linguistics and Psychology*, 31–94. Cham: Springer International Publishing. doi:10.1007/978-3-319-45977-6_3. https://doi.org/10.1007/978-3-319-45977-6_3

Pereira, Fernando. 2000. Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 358(1769). 1239–1253. doi:10.1098/rsta.2000.0583. https://doi.org/10.1098/rsta.2000.0583

Perrigo, Billy. 2023. OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. https://time.com/6247678/openai-chatgpt-kenya-workers/.

Piantadosi, Steven T. forthcoming. Modern language models refute Chomsky's approach to language. In Gibson, Edward & Poliak, Moshe (eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett* (Empirically Oriented Theoretical Morphology and Syntax), Berlin: Language Science Press. https://lingbuzz.net/lingbuzz/007180.

Piantadosi, Steven T. & Tily, Harry & Gibson, Edward. 2012. The communicative function of ambiguity in language. *Cognition* 122(3). 280 – 291. doi:10.1016/j.cognition.2011.10.004. https://doi.org/10.1016/j.cognition.2011.10.004

Pickering, Martin J. & Garrod, Simon. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27(2). 169–190. doi:10.1017/S0140525X04000056. https://doi.org/10.1017/S0140525X04000056

Pickering, Martin J. & Garrod, Simon. 2006. Alignment as the Basis for Successful Communication. *Research on Language and Computation* 4(2). 203–228. doi:10.1007/s11168-006-9004-0. https://doi.org/10.1007/s11168-006-9004-0

Pickering, Martin J. & Garrod, Simon. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences* 36(4). 329–347. doi:10.1017/S0140525X12001495. `https://doi.org/10.1017/S0140525X12001495`

Pickering, Martin J. & Garrod, Simon. 2021. *Understanding Dialogue: Language Use and Social Interaction.* Cambridge, UK: Cambridge University Press. doi:10.1017/9781108610728. `https://doi.org/10.1017/9781108610728`

Pinker, Steven & Bloom, Paul. 1990. Natural language and natural selection. *Behavioral and Brain Sciences* 13(4). 707–727. doi:10.1017/S0140525X00081061. `https://doi.org/10.1017/S0140525X00081061`

Plato. 1997. Cratylus. In Cooper, John M. (ed.), *Plato: Complete Works*, 101–156. Indianapolis, IN: Hackett. `https://n2t.net/ark:/13960/s2wmvgxwttz`.

Powell, Lewis. 2021. Locke on the uses and abuses of language. In Gordon-Roth, Jessica & Weinberg, Shelley (eds.), *The Lockean Mind* (Routledge Philosophical Minds), 181–190. New York and London: Routledge. doi:10.4324/9781315099675. `https://doi.org/10.4324/9781315099675`

Proudfoot, Diane. 2013. Rethinking Turing's Test. *The Journal of Philosophy* 110(7). 391–411. `https://www.jstor.org/stable/43820781`.

Proudfoot, Diane. 2017a. Child machines. In Copeland, Jack & Bowen, Jonathan & Sprevak, Mark & Wilson, Robin (eds.), *The Turing Guide*, 315–325. Oxford: Oxford University Press. doi:10.1093/oso/9780198747826.003.0040. `https://doi.org/10.1093/oso/9780198747826.003.0040`

Proudfoot, Diane. 2017b. Turing's concept of intelligence. In Copeland, Jack & Bowen, Jonathan & Sprevak, Mark & Wilson, Robin (eds.), *The Turing Guide*, 301–307. Oxford: Oxford University Press. doi:10.1093/oso/9780198747826.003.0038. `https://doi.org/10.1093/oso/9780198747826.003.0038`

Proudfoot, Diane. 2020. Rethinking Turing's Test and the Philosophical Implications. *Minds and Machines* 30(4). 487–512. doi:10.1007/s11023-020-09534-7. `https://doi.org/10.1007/s11023-020-09534-7`

Putnam, Hilary. 1975. The Meaning of 'Meaning'. In *Mind, Language and Reality: Philosophical Papers*, vol. 2, 215–271. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511625251.014. `https://doi.org/10.1017/CBO9780511625251.014`

Putnam, Hilary. 1981. *Reason, Truth and History.* Cambridge, UK: Cambridge University Press. `https://n2t.net/ark:/13960/t3dz0rg0c`.

Quetelet, Adolphe. 1835/2013. *A Treatise on Man and the Development of his Faculties.* Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139864909. `https://doi.org/10.1017/CBO9781139864909`

Quetelet, Adolphe. 1846. *Lettres à S. A. R. Le Duc Régnant de Saxe Cobourg et Gotha, sur la théorie des probabilités, appliquée aux sciences morales et politique*. Bruxelles: Hayez. doi:10.3931/e-rara-60003. https://doi.org/10.3931/e-rara-60003

Quine, Willard Van Orman. 1990/1992. *Pursuit of Truth*. Cambridge, MA: Harvard University Press revised edn.

Rad, Mostafa Salari & Martingano, Alison Jane & Ginges, Jeremy. 2018. Toward a psychology of *Homo sapiens*: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences* 115(45). 11401–11405. doi:10.1073/pnas.1721165115. https://doi.org/10.1073/pnas.1721165115

Radford, Alec & Wu, Jeffrey & Child, Rewon & Luan, David & Amodei, Dario & Sutskever, Ilya. 2019. Language Models are Unsupervised Multitask Learners. https://openai.com/blog/better-language-models/.

Ranalli, Brent. 2019. Thoreau's Indian Stride. *The Concord Saunterer* 27. 89–110. https://www.jstor.org/stable/45271429.

Reboul, Anne. 2017. *Cognition and Communication in the Evolution of Language*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780198747314.001.0001. https://doi.org/10.1093/acprof:oso/9780198747314.001.0001

Recanati, François. 2004. *Literal Meaning*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511615382. https://doi.org/10.1017/CBO9780511615382

Reddy, Michael J. 1979/1993. The conduit metaphor: A case of frame conflict in our language about language. In Ortony, Andrew (ed.), *Metaphor and Thought*, 164–201. Cambridge, UK: Cambridge University Press 2nd edn. doi:10.1017/CBO9781139173865.012. https://doi.org/10.1017/CBO9781139173865.012

Rescorla, Michael. 2019. Convention. In Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University summer 2019 edn. https://plato.stanford.edu/archives/sum2019/entries/convention/.

Rescorla, Michael. 2020. The Computational Theory of Mind. In Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Fall 2020 edn. https://plato.stanford.edu/archives/fall2020/entries/computational-mind/.

Rivas, Esteban. 2005. Recent use of signs by chimpanzees (pan troglodytes) in interactions with humans. *Journal of Comparative Psychology* 119(4). 404–417. doi:10.1037/0735-7036.119.4.404. https://doi.org/10.1037/0735-7036.119.4.404

Rodd, Jennifer M. 2018. Lexical Ambiguity. In Rueschemeyer, Shirley-Ann & Gaskell, M. Gareth (eds.), *The Oxford Handbook of Psycholinguistics*, 96–117. Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780198786825.013.5. https://doi.org/10.1093/oxfordhb/9780198786825.013.5

Rodd, Jennifer M. 2020. Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access. *Perspectives on Psychological Science* 15(2). 411–427. doi: 10.1177/1745691619885860. https://doi.org/10.1177/1745691619885860

Rogers, Anna. 2021. Changing the World by Changing the Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2182–2194. Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.170. https://doi.org/10.18653/v1/2021.acl-long.170

Roland, Alex & Shiman, Philip. 2002. *Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993*. Cambridge, MA: MIT Press. https://n2t.net/ark:/13960/s2xddzm64vk.

Rosch, Eleanor. 1973. On the internal structure of perceptual and semantic categories. In Moore, Timothy E. (ed.), *Cognitive Development and Acquisition of Language*, 111–144. San Diego, CA: Academic Press. doi:10.1016/B978-0-12-505850-6.50010-4. https://doi.org/10.1016/B978-0-12-505850-6.50010-4

Rosch, Eleanor. 1975. Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General* 104(3). 192–233. doi:10.1037/0096-3445.104.3.192. https://doi.org/10.1037/0096-3445.104.3.192

Rosch, Eleanor. 1978. Principles of Categorization. In Rosch, Eleanor & Lloyd, Barbara (eds.), *Cognition and Categorization*, 27–48. Hillsdale, NJ: Lawrence Elbaum Associates.

Rose, Todd. 2016. *The End of Average: How We Succeed in a World That Values Sameness*. New York, NY: HarperOne. https://n2t.net/ark:/13960/s2gtbrch0dd.

Ross, John R. 1979. Where's English? In Fillmore, Charles J. & Kempler, Daniel & Wang, William S-Y (eds.), *Individual Differences in Language Ability and Language Behavior*, 127–163. New York, NY: Academic Press. doi:10.1016/B978-0-12-255950-1.50014-7. https://doi.org/10.1016/B978-0-12-255950-1.50014-7

Rousseau, Jean-Jacques. 1754/1984. *A Discourse on Inequality*. London: Penguin.

Rubenstein, Herbert & Goodenough, John B. 1965. Contextual Correlates of Synonymy. *Communications of the ACM* 8(10). 627–633. doi:10.1145/365628.365657. https://doi.acm.org/10.1145/365628.365657

Russell, Stuart J. & Norvig, Peter. 2020. *Artificial Intelligence: A Modern Approach*. Hoboken, NJ: Pearson Education, Inc. 4th edn.

Sahlgren, Magnus. 2008. The distributional hypothesis. *Rivista di Linguistica (Italian Journal of Linguistics)* 20(1). 33–53. https://urn.kb.se/resolve?urn=urn:nbn:se:ri:diva-22393.

Santana, Carlos. 2014. Ambiguity in Cooperative Signaling. *Philosophy of Science* 81(3). 398–422. doi:10.1086/676652. https://doi.org/10.1086/676652

Santana, Carlos. 2016. What Is Language? *Ergo, an Open Access Journal of Philosophy* 3(20201214). doi:10.3998/ergo.12405314.0003.019. https://doi.org/10.3998/ergo.12405314.0003.019

Sassoli de Bianchi, Massimiliano. 2013. The Observer Effect. *Foundations of Science* 18(2). 213–243. doi:10.1007/s10699-012-9298-3. https://doi.org/10.1007/s10699-012-9298-3

Saussure, Ferdinand de. 1916/1971. *Cours de linguistique générale*. Paris: Payot 3rd edn. https://gallica.bnf.fr/ark:/12148/bpt6k314842j.

Saussure, Ferdinand de. 1916/2011. Course in General Linguistics. In Meisel, Perry & Saussy, Haun (eds.), *Course in General Linguistics: Ferdinand de Saussure*, 1–232. New York, NY: Columbia University Press.

Savitsky, Kenneth & Keysar, Boaz & Epley, Nicholas & Carter, Travis & Swanson, Ashley. 2011. The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology* 47(1). 269–273. doi:10.1016/j.jesp.2010.09.005. https://doi.org/10.1016/j.jesp.2010.09.005

Savitzky, Abraham. & Golay, Marcel J. E. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36(8). 1627–1639. doi:10.1021/ac60214a047. https://doi.org/10.1021/ac60214a047

Schafer, Amy J. & Speer, Shari R. & Warren, Paul & White, S. David. 2000. Intonational Disambiguation in Sentence Production and Comprehension. *Journal of Psycholinguistic Research* 29(2). 169–182. doi:10.1023/A:1005192911512. https://doi.org/10.1023/A:1005192911512

Schelling, Thomas C. 1960/1965. *The Strategy of Conflict*. Oxford: Oxford University Press. https://n2t.net/ark:/13960/t80k39v1d.

Schiffrin, Deborah. 1994. *Approaches to Discourse*. Oxford: Blackwell. https://n2t.net/ark:/13960/t0ps77t83.

Scholz, Barbara C. & Pelletier, Francis Jeffry & Pullum, Geoffrey K. & Nefdt, Ryan. 2022. Philosophy of Linguistics. In Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Spring 2022 edn. https://plato.stanford.edu/archives/spr2022/entries/linguistics/.

Scholz, Barbara C. & Pelletier, Francis Jeffry & Pullum, Geoffrey K. & Nefdt, Ryan. 2024. Philosophy of Linguistics. In Zalta, Edward N. & Nodelman, Uri (eds.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Spring 2024 edn. https://plato.stanford.edu/archives/spr2024/entries/linguistics/.

Schutt, Russell K. 2012/2020. Quantitative Methods. In Ritzer, George & Wiedenhoft Murphy, Wendy (eds.), *The Wiley Blackwell Companion to Sociology*, 39–56. Oxford: Wiley Blackwell 2nd edn. doi:10.1002/9781119429333.ch3. https://doi.org/10.1002/9781119429333.ch3

Searle, John R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3). 417–424. doi:10.1017/S0140525X00005756. https://doi.org/10.1017/S0140525X00005756

Searle, John R. 2009. The Turing Test: 55 Years Later. In Epstein, Robert & Roberts, Gary & Beber, Grace (eds.), *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, 139–150. Dordrecht: Springer. doi:10.1007/978-1-4020-6710-5_10. https://doi.org/10.1007/978-1-4020-6710-5_10

Sehlapelo, Martin & Terre Blanche, Martin. 1996. Psychometric Testing in South Africa: View from Above and Below. *Psychology in Society* 21. 49–59.

Shah, Huma & Henry, Odette. 2005. Confederate effect in human–machine textual interaction. In *Proceedings of 5th WSEAS International Conference on Information Science, Communications and Applications (WSEAS ISCA)*. 109–114. Cancun, Mexico: WSEAS.

Shah, Huma & Warwick, Kevin. 2010a. Hidden Interlocutor Misidentification in Practical Turing Tests. *Minds and Machines* 20(3). 441–454. doi:10.1007/s11023-010-9219-6. https://doi.org/10.1007/s11023-010-9219-6

Shah, Huma & Warwick, Kevin. 2010b. Testing Turing's five minutes, parallel-paired imitation game. *Kybernetes* 39(3). 449–465. doi:10.1108/03684921011036178. https://doi.org/10.1108/03684921011036178

Shah, Huma & Warwick, Kevin & Bland, Ian & Chapman, Chris & Allen, Marc. 2012. Turing's Imitation Game: Role of Error-making in Intelligent Thought. In *Turing in Context II: Historical and Contemporary Research in Logic, Computing Machinery and AI*, Gent: Royal Flemish Academy of Belgium for Science and the Arts.

Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27(3). 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shannon, Claude E. 1951. Prediction and Entropy of Printed English. *Bell System Technical Journal* 30(1). 50–64. doi:10.1002/j.1538-7305.1951.tb01366.x. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

Shannon, Claude E. & McCarthy, John. 1956. Preface. In Shannon, Claude E. & McCarthy, John (eds.), *Automata Studies* (Annals of Mathematics Studies, Volume 34), Princeton, NJ: Princeton University Press. https://www.jstor.org/stable/j.ctt1bgzb3s.2.

Shannon, Claude E. & Weaver, Warren. 1949/1964. *The Mathematical Theory of Communication.* Urbana, IL: The University of Illinois Press. https://n2t.net/ark:/13960/t80k66v0t.

Sheynin, Oscar B. 1973. Mathematical Treatment of Astronomical Observations (A Historical Essay). *Archive for History of Exact Sciences* 11(2/3). 97–126. doi:10.1007/BF00343531. https://doi.org/10.1007/BF00343531

Shieber, Stuart M. 2004a. Introduction. In Shieber, Stuart M. (ed.), *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, 1–13. Cambridge, MA: MIT Press. doi:10.7551/mitpress/6928.003.0002. https://doi.org/10.7551/mitpress/6928.003.0002

Shieber, Stuart M. 2004b. The *Bête Machine*. In Shieber, Stuart M. (ed.), *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, 17–21. Cambridge, MA: MIT Press. doi:10.7551/mitpress/6928.003.0004. https://doi.org/10.7551/mitpress/6928.003.0004

Sidnell, Jack. 2014. The architecture of intersubjectivity revisited. In Enfield, Nick J. & Kockelman, Paul & Sidnell, Jack (eds.), *The Cambridge Handbook of Linguistic Anthropology* (Cambridge Handbooks in Language and Linguistics), 364–399. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139342872.018. https://doi.org/10.1017/CBO9781139342872.018

Simpson, Thomas. 1755. A letter to the Right Honourable George Earl of Macclesfield, President of the Royal Society, on the advantage of taking the mean of a number of observations, in practical Astronomy. *Philosophical Transactions of the Royal Society of London* 49. 82–93. doi:10.1098/rstl.1755.0020. https://doi.org/10.1098/rstl.1755.0020

Singer, Judy. 1999. Why can't you be normal for once in your life? From a 'problem with no name' to the emergence of a new category of difference. In Corker, Mairian & French, Sally (eds.), *Disability Discourse* (Disability, Human Rights and Society), 59–67. Buckingham, VA: Open University Press. https://n2t.net/ark:/13960/s2sxh9vpsw2.

Sinha, Ravi & Mihalcea, Rada. 2014. Explorations in lexical sample and all-words lexical substitution. *Natural Language Engineering* 20(1). 99–129. doi:10.1017/S1351324912000265. https://doi.org/10.1017/S1351324912000265

Smith, Neil. 2000. Foreword. In *New Horizons in the Study of Language and Mind*, vi–xvi. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511811937. https://doi.org/10.1017/CBO9780511811937

Smith, Neil & Tsimpli, Ianthi Maria. 1991. Linguistic modularity? A case study of a 'Savant' linguist. *Lingua* 84(4). 315–351. doi:10.1016/0024-3841(91)90034-3. https://doi.org/10.1016/0024-3841(91)90034-3

Sober, Elliott. 1993/2000. *Philosophy of Biology*. Boulder, CO: Westview Press 2nd edn.

Sparck Jones, Karen & Galliers, Julia R. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Berlin and Heidelberg: Springer-Verlag. doi:10.1007/bfb0027470. https://doi.org/10.1007/bfb0027470

Sperber, Dan & Cara, Francesco & Girotto, Vittorio. 1995. Relevance theory explains the selection task. *Cognition* 57(1). 31–95. doi:10.1016/0010-0277(95)00666-M. https://doi.org/10.1016/0010-0277(95)00666-M

Sperber, Dan & Origgi, Gloria. 2010. A pragmatic perspective on the evolution of language. In Larson, Richard K. & Déprez, Viviane & Yamakido, Hiroko (eds.), *The Evolution of Human Language: Biolinguistic Perspectives* (Approaches to the Evolution of Language), 124–132. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511817755.009. https://doi.org/10.1017/CBO9780511817755.009

Sperber, Dan & Wilson, Deirdre. 1986/1995. *Relevance: Communication and Cognition*. Oxford: Basil Blackwell 2nd edn.

Sperber, Dan & Wilson, Deirdre. 1998. The Mapping between the Mental and the Public Lexicon. In Peter Carruthers, Jill Boucher (ed.), *Language and Thought: Interdisciplinary Themes*, 184–200. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511597909.012. https://doi.org/10.1017/CBO9780511597909.012

Sperber, Dan & Wilson, Deirdre. 2012. Introduction: pragmatics. In Wilson, Deirdre & Sperber, Dan (eds.), *Meaning and Relevance*, 1–28. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139028370.002. https://doi.org/10.1017/CBO9781139028370.002

Sprouse, Jon & Almeida, Diogo. 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics* 48(3). 609–652. doi:10.1017/S0022226712000011. https://doi.org/10.1017/S0022226712000011

Sprouse, Jon & Almeida, Diogo. 2013. The role of experimental syntax in an integrated cognitive science of language. In Boeckx, Cedric & Grohmann, Kleanthes K. (eds.), *The Cambridge Handbook of Biolinguistics* (Cambridge Handbooks in Language and Linguistics), 181–202. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511980435.013. https://doi.org/10.1017/CBO9780511980435.013

Sprouse, Jon & Almeida, Diogo. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa* 2(1). 1. doi:10.5334/gjgl.236. https://doi.org/10.5334/gjgl.236/

Sprouse, Jon & Caponigro, Ivano & Greco, Ciro & Cecchetto, Carlo. 2016. Experimental syntax and the variation of island effects in English and Italian. *Natural Language & Linguistic Theory* 34(1). 307–344. doi:10.1007/s11049-015-9286-8. https://doi.org/10.1007/s11049-015-9286-8

Sprouse, Jon & Schütze, Carson T. & Almeida, Diogo. 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua* 134. 219–248. doi:10.1016/j.lingua.2013.07.002. https://doi.org/10.1016/j.lingua.2013.07.002

Sprouse, Jon & Yankama, Beracah & Indurkhya, Sagar & Fong, Sandiway & Berwick, Robert C. 2018. Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review* 35(3). 575–599. doi:10.1515/tlr-2018-0005. https://doi.org/10.1515/tlr-2018-0005

Stahl, Saul. 2006. The Evolution of the Normal Distribution. *Mathematics Magazine* 79(2). 96–113. doi:10.1080/0025570X.2006.11953386. https://doi.org/10.1080/0025570X.2006.11953386

Stainton, Robert J. 2011. In defense of public languages. *Linguistics and Philosophy* 34(5). 479–488. doi:10.1007/s10988-011-9104-7. https://doi.org/10.1007/s10988-011-9104-7

Stainton, Robert J. 2016. A Deranged Argument Against Public Languages. *Inquiry* 59(1). 6–32. doi:10.1080/0020174X.2015.1115279. https://doi.org/10.1080/0020174X.2015.1115279

Stede, Manfred. 2012. *Discourse Processing* (Synthesis Lectures on Human Language Technologies). San Rafael, CA: Morgan & Claypool. doi:10.1007/978-3-031-02144-2. https://doi.org/10.1007/978-3-031-02144-2

Sternberg, Robert J. 1984. Toward a triarchic theory of human intelligence. *Behavioral and Brain Sciences* 7(2). 269–287. doi:10.1017/S0140525X00044629. https://doi.org/10.1017/S0140525X00044629

Sternberg, Robert J. 1985. *Beyond IQ: A triarchic theory of human intelligence*. Cambridge, UK: Cambridge University Press. https://n2t.net/ark:/13960/t19m41v06.

Sternberg, Robert J. 2004. Culture and Intelligence. *American Psychologist* 59(5). 325–338. doi:10.1037/0003-066X.59.5.325. https://doi.org/10.1037/0003-066X.59.5.325

Sternberg, Robert J. (ed.). 2020. *The Cambridge Handbook of Intelligence*. Cambridge, UK: Cambridge University Press. doi:10.1017/9781108770422. https://doi.org/10.1017/9781108770422

Stiennon, Nisan & Ouyang, Long & Wu, Jeffrey & Ziegler, Daniel & Lowe, Ryan & Voss, Chelsea & Radford, Alec & Amodei, Dario & Christiano, Paul F. 2020. Learning to summarize with human feedback. In Larochelle, H. & Ranzato, M. & Hadsell, R. & Balcan, M. F. & Lin, H. (eds.), *Advances in Neural Information Processing Systems*, vol. 33. 3008–3021. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Stiff, Chris. 2012. Watch What You Write: How Errors in Feedback Influence Consumer Attitudes and Behavior. *Journal of Internet Commerce* 11(1). 41–67. doi:10.1080/15332861.2012.650988. https://doi.org/10.1080/15332861.2012.650988

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: Harvard University Press. https://n2t.net/ark:/13960/t2h79116k.

Streeter, Lynn A. 1978. Acoustic determinants of phrase boundary perception. *The Journal of the Acoustical Society of America* 64(6). 1582–1592. doi:10.1121/1.382142. https://doi.org/10.1121/1.382142

Swets, Benjamin & Desmet, Timothy & Clifton, Charles & Ferreira, Fernanda. 2008. Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition* 36(1). 201–216. doi:10.3758/MC.36.1.201. https://doi.org/10.3758/MC.36.1.201

Szabó, Zoltán Gendler. 2022. Compositionality. In Zalta, Edward N. & Nodelman, Uri (eds.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Fall 2022 edn. https://plato.stanford.edu/archives/fall2022/entries/compositionality/.

Tafreshi, Donna. 2022. Adolphe Quetelet and the Legacy of the “Average Man” in Psychology. *History of Psychology* 25(1). 34–55. doi:10.1037/hop0000202. https://doi.org/10.1037/hop0000202

Taylor, Talbot J. 1992. *Mutual Misunderstanding: Scepticism and the Theorizing of Language and Interpretation*. Durham, NC: Duke University Press. doi:10.1515/9780822383000. https://doi.org/10.1515/9780822383000

Taylor, Talbot J. & Cameron, Deborah. 1987. *Analysing Conversation: Rules and Units in the Structure of Talk*. Oxford: Pergamon Press. https://n2t.net/ark:/13960/t0hv0b71t.

The Economist. 1992. Artificial Stupidity. *The Economist* 324(7770). https://www.caiac.ca/sites/default/files/shared/canai-archives/CAI%20Volume%2032%20-%20Summer%201993.pdf.

Thompson, D’Arcy Wentworth. 1917/1992. *On Growth and Form*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781107325852. https://doi.org/10.1017/CBO9781107325852

Thurstone, Louis L. 1938. *Primary Mental Abilities*. Chicago, IL: The University of Chicago Press. https://n2t.net/ark:/13960/s27t6cgm9mx.

Tilley, Helen. 2011. *Africa as a Living Laboratory: Empire, Development, and the Problem of Scientific Knowledge, 1870-1950*. Chicago, IL: University of Chicago

Press. doi:10.7208/chicago/9780226803487.001.0001. https://doi.org/10.7208/chicago/9780226803487.001.0001

Tomasello, Michael. 1999. *The Cultural Origins of Human Cognition*. Cambridge, MA: MIT Press. https://n2t.net/ark:/13960/s2vvqmm5mrt.

Tomasello, Michael. 2008. *Origins of Human Communication*. Cambridge, MA: MIT Press. https://n2t.net/ark:/13960/s2cqx8mz0bd.

Tomasello, Michael. 2014. *A Natural History of Human Thinking*. Cambridge, MA: MIT Press. doi:10.4159/9780674726369. https://doi.org/10.4159/9780674726369

Tomasello, Michael. 2019. *Becoming Human: A Theory of Ontogeny*. Cambridge, MA: Harvard University Press. doi:10.4159/9780674988651. https://doi.org/10.4159/9780674988651

Tomasello, Michael & Kruger, Ann Cale & Ratner, Hilary Horn. 1993. Cultural learning. *Behavioral and Brain Sciences* 16(3). 495–511. doi:10.1017/S0140525X0003123X. https://doi.org/10.1017/S0140525X0003123X

Traxler, Matthew J. & Pickering, Martin J. & Clifton, Charles. 1998. Adjunct Attachment Is Not a Form of Lexical Ambiguity Resolution. *Journal of Memory and Language* 39(4). 558–592. doi:10.1006/jmla.1998.2600. https://doi.org/10.1006/jmla.1998.2600

Trott, Sean & Bergen, Benjamin. 2022. Languages are efficient, but for whom? *Cognition* 225. 105094. doi:10.1016/j.cognition.2022.105094. https://doi.org/10.1016/j.cognition.2022.105094

Turing, Alan. 1947/2004. Lecture on the Automatic Computing Engine. In Copeland, B. Jack (ed.), *The Essential Turing*, 362–394. Oxford: Oxford University Press. doi:10.1093/oso/9780198250791.003.0015. https://doi.org/10.1093/oso/9780198250791.003.0015

Turing, Alan. 1948/2004. Intelligent Machinery. In Copeland, B. Jack (ed.), *The Essential Turing*, 410–432. Oxford: Oxford University Press. doi:10.1093/oso/9780198250791.003.0016. https://doi.org/10.1093/oso/9780198250791.003.0016

Turing, Alan & Braithwaite, Richard & Jefferson, Geoffrey & Newman, Max. 1952/2004. Can Automatic Calculating Machines Be Said To Think? In Copeland, B. Jack (ed.), *The Essential Turing*, 494–506. Oxford: Oxford University Press. doi:10.1093/oso/9780198250791.003.0020. https://doi.org/10.1093/oso/9780198250791.003.0020

Turing, Alan M. 1950. Computing machinery and intelligence. *Mind* LIX(236). 433–460. doi:10.1093/mind/LIX.236.433. https://doi.org/10.1093/mind/LIX.236.433

Turing, Alan M. 1950/2009. Computing Machinery and Intelligence. In Epstein, Robert & Roberts, Gary & Beber, Grace (eds.), *Parsing the Turing Test: Philosophical* *and Methodological Issues in the Quest for the Thinking Computer*, 23–65. Dordrecht: Springer Netherlands. doi:10.1007/978-1-4020-6710-5_3. https://doi.org/10.1007/978-1-4020-6710-5_3

Turner, Mark. 1991. *Reading Minds: The Study of English in the Age of Cognitive Science*. Princeton, NJ: Princeton University Press. doi:10.2307/j.ctv1ddcz96. https://doi.org/10.2307/j.ctv1ddcz96

Turney, Peter D. & Pantel, Patrick. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37(1). 141–188. doi:10.1613/jair.2934. https://doi.org/10.1613/jair.2934

Vanderschraaf, Peter & Sillari, Giacomo. 2022. Common Knowledge. In Zalta, Edward N. & Nodelman, Uri (eds.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University Fall 2022 edn. https://plato.stanford.edu/archives/fall2022/entries/common-knowledge/.

Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan N. & Kaiser, Łukasz & Polosukhin, Illia. 2017. Attention is All you Need. In Guyon, I. & Luxburg, U. Von & Bengio, S. & Wallach, H. & Fergus, R. & Vishwanathan, S. & Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vernon, Philip E. 1965. Ability factors and environmental influences. *American Psychologist* 20(9). 723–733. doi:10.1037/h0021472. https://doi.org/10.1037/h0021472

Vernon, Philip E. 1969/2013. *Intelligence and Cultural Environment* (Psychology Revivals). New York, NY: Routledge. doi:10.4324/9781315879949. https://doi.org/10.4324/9781315879949

Virtanen, Pauli & Gommers, Ralf & Oliphant, Travis E. & Haberland, Matt & Reddy, Tyler & Cournapeau, David & Burovski, Evgeni & Peterson, Pearu & Weckesser, Warren & Bright, Jonathan & van der Walt, Stéfan J. & Brett, Matthew & Wilson, Joshua & Millman, K. Jarrod & Mayorov, Nikolay & Nelson, Andrew R. J. & Jones, Eric & Kern, Robert & Larson, Eric & Carey, C J & Polat, İlhan & Feng, Yu & Moore, Eric W. & VanderPlas, Jake & Laxalde, Denis & Perktold, Josef & Cimrman, Robert & Henriksen, Ian & Quintero, E. A. & Harris, Charles R. & Archibald, Anne M. & Ribeiro, Antônio H. & Pedregosa, Fabian & van Mulbregt, Paul & SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17. 261–272. doi:10.1038/s41592-019-0686-2. https://doi.org/10.1038/s41592-019-0686-2

Wachbroit, Robert. 1994. Normality as a Biological Concept. *Philosophy of Science* 61(4). 579–591. doi:10.1086/289823. https://doi.org/10.1086/289823

Wang, Pei. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2). 1–37. doi:10.2478/jagi-2019-0002. https://doi.org/10.2478/jagi-2019-0002

Wang, Pei & Goertzel, Ben. 2007. Introduction: Aspects of Artificial General Intelligence. In Goertzel, Ben & Wang, Pei (eds.), *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, 1–16. Amsterdam: IOS Press. https://ebooks.iospress.nl/volumearticle/3470.

Warwick, Kevin & Shah, Huma. 2014. Good Machine Performance in Turing's Imitation Game. *IEEE Transactions on Computational Intelligence and AI in Games* 6(3). 289–299. doi:10.1109/TCIAIG.2013.2283538. https://doi.org/10.1109/TCIAIG.2013.2283538

Warwick, Kevin & Shah, Huma. 2015. Human misidentification in Turing tests. *Journal of Experimental & Theoretical Artificial Intelligence* 27(2). 123–135. doi:10.1080/0952813X.2014.921734. https://doi.org/10.1080/0952813X.2014.921734

Warwick, Kevin & Shah, Huma. 2016a. Can machines think? A report on Turing test experiments at the Royal Society. *Journal of Experimental & Theoretical Artificial Intelligence* 28(6). 989–1007. doi:10.1080/0952813X.2015.1055826. https://doi.org/10.1080/0952813X.2015.1055826

Warwick, Kevin & Shah, Huma. 2016b. Effects of lying in practical Turing tests. *AI & SOCIETY* 31(1). 5–15. doi:10.1007/s00146-013-0534-3. https://doi.org/10.1007/s00146-013-0534-3

Wasow, Thomas. 2015. Ambiguity Avoidance is Overrated. In Winkler, Susanne (ed.), *Language and Communication*, 29–48. Berlin: De Gruyter. doi:10.1515/9783110403589-003. https://doi.org/10.1515/9783110403589-003

Wasow, Thomas & Arnold, Jennifer. 2005. Intuitions in linguistic argumentation. *Lingua* 115(11). 1481–1496. doi:10.1016/j.lingua.2004.07.001. https://doi.org/10.1016/j.lingua.2004.07.001

Wesson, Robert G. 1991. *Beyond Natural Selection*. Cambridge, MA: MIT Press. https://n2t.net/ark:/13960/s2wqw3t417k.

Willimans, George C. 1966/1996. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton, NJ: Princeton University Press. https://n2t.net/ark:/13960/t8rc56w90.

Wilson, Deirdre. 2018. Relevance Theory and Literary Interpretation. In Cave, Terence & Wilson, Deirdre (eds.), *Reading Beyond the Code: Literature and Relevance Theory*, 185–204. Oxford: Oxford University Press. doi:10.1093/oso/9780198794776.003.0011. https://doi.org/10.1093/oso/9780198794776.003.0011

Wilson, Deirdre & Sperber, Dan. 2002. Truthfulness and Relevance. *Mind* 111(443). 583–632. doi:10.1093/mind/111.443.583. `https://doi.org/10.1093/mind/111.443.583`

Wilson, Deirdre & Sperber, Dan. 2004/2006. Relevance Theory. In Horn, Laurence R. & Ward, Gregory (eds.), *The Handbook of Pragmatics*, 607–632. Oxford: Blackwell. doi:10.1002/9780470756959.ch27. `https://doi.org/10.1002/9780470756959.ch27`

Wilson, Deirdre & Sperber, Dan (eds.). 2012. *Meaning and Relevance*. Cambridge, UK: Cambridge University Press.

Winch, Peter. 1958/2008. *The Idea of a Social Science and its Relation to Philosophy*. London: Routledge Routledge Classics edn. doi:10.4324/9780203820766. `https://doi.org/10.4324/9780203820766`

Winograd, Terry. 1972. Understanding Natural Language. *Cognitive Psychology* 3(1). 1–191. doi:10.1016/0010-0285(72)90002-3. `https://doi.org/10.1016/0010-0285(72)90002-3`

Winograd, Terry. 1980. What Does it Mean to Understand Language? *Cognitive Science* 4(3). 209–241. doi:10.1207/s15516709cog0403_1. `https://doi.org/10.1207/s15516709cog0403_1`

Wittgenstein, Ludwig. 1953/2009. *Philosophical Investigations*. Oxford: Wiley-Blackwell 4th edn.

Wittgenstein, Ludwig. 1974. *Philosophical Grammar*. Oxford: Basil Blackwell.

Zamzow, Rachel. 2023. Autism researchers face off over language. *Science* 379(6632). 523–524. doi:10.1126/science.adh0585. `https://doi.org/10.1126/science.adh0585`

Zhou, Xuhui & Sap, Maarten & Swayamdipta, Swabha & Choi, Yejin & Smith, Noah. 2021. Challenges in Automated Debiasing for Toxic Language Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 3143–3155. Online: Association for Computational Linguistics. `https://www.aclweb.org/anthology/2021.eacl-main.274`.

Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. New York, NY: Addison-Wesley. `https://n2t.net/ark:/13960/t4xh50h6z`.

Zlatev, Jordan & Racine, Timothy P. & Sinha, Chris & Itkonen, Esa (eds.). 2008. *The Shared Mind: Perspectives on intersubjectivity*. Amsterdam: John Benjamins. doi:10.1075/celcr.12. `https://doi.org/10.1075/celcr.12`

Zwaan, Rolf A. & Radvansky, Gabriel A. 1998. Situation models in language comprehension and memory. *Psychological Bulletin* 123(2). 162–185. doi:10.1037/0033-2909.123.2.162. `https://doi.org/10.1037/0033-2909.123.2.162`