Artificial neural network models of language are mostly known and appreciated today for providing a backbone for formidable AI technologies. This thesis takes a different perspective. Through a series of studies on language comprehension and production, it investigates whether artificial neural networks—beyond being useful in countless AI applications—can serve as accurate computational simulations of human language use, and thus as a new core methodology for the language sciences.

UNIVERSITY OF AMSTERDAM

Neural Models of Language Use
Mario Giulianelli

Neural Models of Language Use

Studies of Language Comprehension and Production in Context

Mario Giulianelli

UNIVERSITY OF AMSTERDAM

Neural Models of Language Use

Studies of Language Comprehension and Production in Context

Mario Giulianelli

Neural Models of Language Use

Studies of Language Comprehension and Production in Context

ILLC Dissertation Series DS-2023-10

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl

The research for this doctoral thesis has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

Cover design by Mario Giulianelli.
Printed and bound by Ipskamp Printing.

ISBN: 978-94-6473-254-2

Neural Models of Language Use
Studies of Language Comprehension and Production in Context

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op vrijdag 15 december 2023, te 11.00 uur

door Mario Giulianelli
geboren te Rome

Promotiecommissie

Promotor:	prof. dr. R. Fernández Rovira	Universiteit van Amsterdam
Copromotor:	dr. W.H. Zuidema	Universiteit van Amsterdam
Overige leden:	prof. dr. G. Boleda	Universitat Pompeu Fabra
	prof. dr. V. Demberg	Universität des Saarlandes
	dr. I.A. Titov	Universiteit van Amsterdam
	prof. dr. F. Roelofsen	Universiteit van Amsterdam
	dr. S. Pezzelle	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Contents

Acknowledgments **ix**

1 Introduction **1**
1.1 A story of language use 2
1.2 Towards a computational theory of language use 5
1.3 Main contributions and overview 6

2 General background **13**
2.1 Artificial neural networks as models of language use 13
2.2 Language modelling 14

Part One: Word Usage

3 Background **27**
3.1 Word representations in NLP 27
3.2 Diachronic word usage change 30
3.3 Definition modelling 33

4 Contextualised neural word representations **35**
4.1 Introduction 36
4.2 A usage-based approach to lexical semantic change modelling . . . 37
4.3 Data 41
4.4 Correlation with human judgements 41
4.5 Qualitative analysis 47
4.6 Evaluation across languages: The SemEval-2020 shared task . . . 51
4.7 Conclusion 58

5 Contextualised definitions as interpretable word representations 61
5.1 Introduction . . . 62
5.2 Data . . . 63
5.3 Definition generation . . . 64
5.4 Definitions are interpretable word representations . . . 68
5.5 Labelling word senses with definitions . . . 71
5.6 Explaining semantic change with sense labels . . . 74
5.7 Conclusion . . . 77

Part Two: Utterance Comprehension

6 Estimating surprisal with neural language models 83
6.1 Background . . . 84
6.2 Method . . . 85
6.3 Data . . . 87
6.4 Experimental setup . . . 89
6.5 Analysis of language model estimates . . . 90
6.6 Replication study: Surprisal constancy in newspaper articles . . . 92

7 Utterance surprisal as a function of discourse context 95
7.1 Introduction . . . 96
7.2 Surprisal throughout texts and dialogues: Constancy vs. uniformity 97
7.3 Surprisal within contextual units . . . 101
7.4 Discussion and conclusions . . . 109

8 The facilitating effect of construction repetition 113
8.1 Introduction . . . 114
8.2 Constructions . . . 115
8.3 Data: Extracting repeated constructions . . . 117
8.4 Experimental setup . . . 119
8.5 Preliminary experiments . . . 120
8.6 The facilitating effect of construction repetition . . . 123
8.7 Discussion and conclusions . . . 129

Part Three: Utterance Production

9 Background 137
9.1 Natural language generation . . . 137
9.2 Expectations, predictability, and surprisal . . . 139

10 Evaluating uncertainty in neural text generators 141
10.1 Introduction . . . 142
10.2 Probing language processes for production variability . . . 143
10.3 Experimental setup . . . 145
10.4 Human production variability across NLG tasks . . . 147
10.5 Neural text generators' compliance to human production variability 148
10.6 Qualitative instance-level analysis . . . 153
10.7 Discussion and conclusions . . . 155

11 Measuring utterance predictability with neural text generators 159
11.1 Introduction . . . 160
11.2 Alternatives in semantics and pragmatics . . . 161
11.3 Alternative-based information value . . . 161
11.4 Experimental setup . . . 163
11.5 The psychometric predictive power of information value . . . 164
11.6 In-depth analysis of psychometric data . . . 168
11.7 Relation to utterance surprisal . . . 171
11.8 Discussion and conclusions . . . 174

12 Towards human-like production strategies in NLG systems 177
12.1 Introduction . . . 178
12.2 Doing things with words . . . 178
12.3 Case study 1: Reference games . . . 181
12.4 Case study 2: Text summarisation . . . 182
12.5 Pragmatic production strategies . . . 183
12.6 Conclusion . . . 185

13 Conclusion 187

A Word usage 191
A.1 Preliminary analysis of usage examples . . . 191
A.2 Prompt selection . . . 191
A.3 Additional results: Other models and model variants . . . 194
A.4 Additional examples of generated definitions and sense labels . . . 194
A.5 Human evaluation guidelines . . . 196
A.6 Clustering embedding spaces . . . 197

B Utterance comprehension 201
B.1 Corpus excerpts . . . 201
B.2 Language models . . . 201
B.3 Replication study: Surprisal constancy in newspaper articles . . . 205
B.4 Results: Utterance surprisal as a function of discourse context . . . 208
B.5 Extraction of repeated constructions . . . 208

B.6 Adaptive language model 210
B.7 Results: The facilitating effect of construction repetition 213

C Utterance production **217**
C.1 Further figures on production variability 217
C.2 Alternative set generators 220
C.3 Psychometric predictive power and sensitivity of information value 221
C.4 Utterance-level surprisal 221
C.5 Intrinsic robustness analysis 224
C.6 More derived measures of information value 224
C.7 Results: The psychometric predictive power of information value . 225

Samenvatting **265**

Abstract **267**

Acknowledgments

My greatest thanks go to Raquel. *Raquel,* thank you for your patience, for always listening to me, for your support in difficult times. Thank you for introducing me to Janie, for creating a safe and exciting environment to work in, for teaching me how to write and think science. From your example, I learned what it means to be a scientist. I am so lucky to work with you.

I am also very grateful to all the amazing people and researchers in the ILLC Dialogue Modelling Group. I would like to thank, in particular, Sandro, Marco, Ece, and Joris. Our time together and our discussions have been invaluable to me. I don't know if I will ever find another group like ours.

I would also thank Andrey, Wilker, and Dieuwke. I have learned so much from you all. *Andrey,* working with you is always an incredible pleasure.

Janie, thank you for being a great collaborator and a wonderful person. We are the best team.

Giorgio and Fabrizia, thank you for making Amsterdam so fun. *Angelo,* our calls, especially during the pandemic, have kept me (almost) sane—thank you. Thanks to all of my 'Roman friends' for always making me feel like I have never left and for your support, whenever I needed it. I love my job, but surviving a PhD wouldn't have been possible without you all. I promise I love you even more.

Alessandro, Elisabetta, Simone, thank you for always believing in me. You make me feel like I owe you nothing, but I owe you everything. *Grazie.*

Chapter 1

Introduction

Humans use bits of behaviour such as speech, hand and body movements to transmit information among themselves. They use linguistic behaviour to communicate knowledge, plans, emotions, values, and more in general to achieve goals in the world, i.e., to change the state of the environment in which they are situated. For example, humans use language to update the beliefs of other humans, to coordinate their activity, and to complete tasks jointly. Being able to describe how humans learn and exercise this ability is to me the single most exhilarating goal of scientific inquiry.

The ultimate motivation that guides my research—a good portion of which, as of today, is described in this thesis—is reverse engineering the human ability to exchange information through linguistic behaviour. I believe achieving this goal coincides with the creation of a *model* that can use language *like* and *with* humans and serve as a computational theory of language use. Insights from this line of research and progress towards this ideal model, a controllable but fully autonomous contextually-grounded language agent, will inform and stimulate the development of language technologies that more faithfully reproduce human linguistic behaviour.

The road towards this goal is long and the life of a PhD student too short to walk it. But it is with a feeling of total bliss that I can say that I now have a map and an approximate route. Looking back—and looking forward—this is to me the most exciting and valuable outcome of my doctoral studies. Be that as it may, becoming and being a scientist is a form of service, and the community I am serving, whether widely or more narrowly delineated, is fully indifferent to my internal struggles, tensions and learning outcomes—and rightly so. The map may be imperfect, the route is certainly provisional, and the steps I took were sometimes wobbly, but they are my contribution to the 'community' and I will proudly outline them in this thesis. I hope to be able to look back to these pages with a refined route, more steps to count, and the same feeling of fulfilment.

1.1 A story of language use

Language use is a type of behaviour, and just like any type of behaviour, it is embedded in an environment, the *context* of language use. The context, with its affordances, determines which actions can be taken; it shapes and constrains what things can be said and how they can be said. A crucial component of the environment in which humans take actions are other humans. Humans certainly talk to themselves, or use language to reason about possible states of the world and to form ideas (those, too, after all, are part of the state of the world; they are physical connections and activations in the human brain) but the primary form of language use is in interaction with other humans.

The reason why humans take actions is to change the state of the environment. In particular, we try to change the state of the environment towards new states that either coincide with or get us closer to our *goals*. Imagine a room containing a desk, a chair, a sofa, a window, and two humans, Peter and Sally. Peter is at the desk and Sally is sitting on the sofa, closer to the window. The window is open and Peter is cold and tired. He thinks that if the window was closed, he would be warmer and happier. Now, with the goal to change the state of the environment—in particular, the state of the room—into one in which the window is closed, Peter may ask Sally 'Could you please close the window?'.

Goals are transformations of the state of the environment into states with favourable properties: for example, states in which we are happier, experience more pleasure, or make less effort. More generally, goals are states which produce positive social, cognitive, and physical effects. The positive effects that states of the environment generate are their *utility*. Peter is really cold, and asking Sally to close the window may have an immediate positive physical effect: Sally might close the window and, soon after, Peter would experience an increase in body temperature. A body temperature increase is the utility of this possible future state of the room. Actions, by extension, also have or generate utility, namely the utility corresponding to the environment states that result from taking those actions. If asking 'Could you please close the window?' transforms the room into a state in which the window is closed, this will in turn generate a temperature increase and positive effects on Peter's body.

Utility can be both positive and negative. Certain states of the world may have positive social effects but negative cognitive effects. For example, being kind and polite in interactions tends to have positive social effects—e.g., it makes us nicer people to talk or it can make us more convincing—but at the same time, more often than not, it requires more physical and cognitive effort. Peter is tired, and utters 'Close the window', which is a much less effortful piece of behaviour to produce. In other words, the negative utility of 'Close the window'—in this case, its physical and cognitive *cost*—is lower in comparison to that of 'Could you please close the window?'.

The interlocutor, or *audience*—here, Sally—perceives the state of the environ-

ment and the bit of linguistic behaviour produced by the speaker—Peter. Relying on her model of the environment and her ability to recognise other humans' goals and plans (via models of the speaker and their utility), the audience uses the speaker's behaviour as a set of instructions to reconstruct, or predict, the new state of the environment that the speaker intended to communicate. Linguistic interaction is successful when the audience's reconstruction of the speaker's goal is faithful to the originally intended new state of the environment. The act of comprehending an utterance also comes with efforts and positive utility. For example, the audience must pay attention to the speaker's behaviour, it must take some time to process it, and it must interpret it within the environment. For all of these actions, the audience pays some cognitive cost, which may be recompensed, for exampl, by positive social utility. If Sally takes the time and effort to comprehend Peter's request, and decides to indeed open the window, she might make everyone happier.

The audience may be physically co-present with the speaker as in the case of Peter and Sally, they might only perceive the linguistic act after one hundred years as it is the case with books, they might be purely hypothetical, or they might be the speaker themselves. In any case, a linguistic act is inseparable from its audience, just like it is inseparable from its context(s) of production and comprehension. This is as true for language production in face-to-face spoken dialogue as it is for a theatre monologue, a podcast, or a written text. Because the connection between linguistic behaviour and underlying communicative intent is purely arbitrary, and because intents are non-observable (at least until new groundbreaking scientific discoveries), a linguistic act is only complete when it is perceived and interpreted by an audience. In fact, if it is perceived multiple times, potentially by different audiences—song lyrics are an example of this—the same linguistic act can induce multiple changes of the state of the world. These changes can be different each time, they might vary for different audiences, and there is no guarantee that the new state of the world will be the one the speaker intended.

The reason why linguistic acts change the state of the world—that is, beyond their mere occurrence, which itself is an obvious world change—is that while linguistic behaviour is an arbitrary encoding of communicative intents, groups of humans converge on similar models of the world and develop shared systems of interpretation. They interpret the same behaviours in similar ways.

Humans continuously make and accept proposals about how language should be used, and the most successful of these proposals become conventionalised. Peter and Sally's is a story about a single interaction, in a single environment, between two individuals. These are in a sense units on the temporal, spacial, and social dimensions of context that constitute the space of language interaction. The temporal, spacial, and social dimension jointly shape and constrain what things can be said and how they can be said. When we combine these three dimensions and observe them at once, we appreciate *language* as a complex adaptive system

of interactions. The relationships between the system's parts, i.e., among humans and between humans and the environment, give rise to ever-changing collective forms of linguistic behaviour.

What makes the emergence of such system of interactions possible are a few fundamental characteristics of linguistic behaviour—so to say, a few preconditions. The first condition is agency, i.e., the ability to behave non-randomly but rather as a function of one's utility and model of the environment. Utility generates intents; the model of the environment prescribes what actions are more likely to lead to intended outcomes and what outcomes are plausible given the current environment state. The ability to speak a certain natural language can be subsumed under this model of the environment. Moreover, because other humans are part of the environment, the model of the environment is also a model of interlocutors. Without the socio-cognitive skill to entertain a model of the interlocutor, language interaction is hardly ever successful, and language cannot be learned in the first place.

The model of the interlocutor—whether a speaker or a comprehender—must prescribe (i) that the interlocutor's actions can be assumed to have an intended utility, and (ii) that the interlocutor makes the same assumptions. If Sally did not ascribe any intent to Peter, she would not open the window—nor would she even try to interpret Peter's utterance as a request. She would rather take it as an arbitrary piece of behaviour to observe, but with which she has nothing to do, like a thunder or a falling leaf. If Peter did not assume that Sally believes he has intents, there would be no point in requesting anything from her. This mutual and recursive recognition of agency (i.e., entertaining intents and performing behaviour as a function of those intents) is a prerequisite for joint action, and thus a second necessary condition for language interaction to occur.

To predict their *joint utility*, humans use mental models, but choosing which utterance to produce and how to interpret it towards joint utility maximisation are decision-making problems. Beyond models of the environment and interlocutors, speakers must thus possess a *higher-level model of utility*, either implicit or explicit, as a way to determine the relative importance of different lower-level utilities and to modulate producer and comprehender's utilities; as well as the *ability to deal with uncertainty*: as humans do not possess a perfect model of the world, they cannot be certain about the transformations of the state of the world their actions will cause. If Peter had been wise, when choosing between the two alternative utterances above, he would have considered that not only does 'Close the window' have lower or negative social utility but also (and probably as a consequence) it creates more uncertainty about the future state of the environment. If he had been wise, he would have thought of the last times he and Sally worked in that room with the window open. A few times he used the utterance 'Close the window' and this resulted in a state of the room in which Sally told him to ask more politely, and a few more linguistic actions were needed to bring back their joint social utility to positive levels. Peter, however, is tired, and these

considerations require an amount of cognitive effort that he is not in a position to expend. So he says 'Close the window' and, more or less knowingly, he takes a higher risk.

Luckily, Sally has an easier time making social utility calculations, she has a good model of Peter, and she knows he is tired. She opens the window anyways.

1.2 Towards a computational theory of language use

This story of language use is a synthesis of multiple philosophical perspectives, scientific theories, and experimental traditions. It is rooted in semiotics (Peirce, 1932; Wittgenstein, 1953; de Saussure, 1972), formal pragmatics (Searle, 1969; Grice, 1975; Sperber and Wilson, 1986; Levinson et al., 2000; Horn and Ward, 2004), usage-based accounts of natural language (Clark, 1996; Tomasello, 2003; Goldberg, 2006; Bybee, 2006, 2010), conversation analysis (Sacks, 1992; Schegloff, 1992; Seedhouse, 2013), rational speech act theory (Frank and Goodman, 2012b; Goodman and Frank, 2016), the concept of situation models (Johnson-Laird, 1986; Sanford and Garrod, 1981; Van Dijk et al., 1983; Zwaan and Radvansky, 1998) and theories of their interactive alignment (Pickering and Garrod, 2004; Garrod and Anderson, 1987).

In addition to synthesising these traditions, my thesis is an attempt to integrate them with a computational modelling approach, that of connectionism and parallel distributed processing (PDP; McClelland et al., 1986; Rumelhart and McClelland, 1986; Lake et al., 2017; McClelland et al., 2019). This is a computational modelling approach which explores how cognitive processes emerge from the interactions among simple, neuron-like units through their weighted connections. The peculiarity of PDP models, which we will refer to as artificial neural network models—or, in short, *neural models*—is that they employ distributed representations. That is, their representation of an item is distributed across multiple neural units (which also participate in representing other items) and their information processing requires the collective and simultaneous involvement of multiple units (McClelland and Rogers, 2003).

While their processing and reasoning mechanisms are different from humans', artificial neural networks can predict human behaviour exceptionally well. Neural models of language, in particular, are powerful models of human language comprehension and production—they can infer word and utterance meaning in context, form human-like expectations about upcoming linguistic material, and generate language which is hardly distinguishable from human language productions. What is crucial is that to do so, they require only a few assumptions: that the knowledge which governs processing is stored in the strengths of a model's connections and that it is acquired gradually through experience. Moreover, neural models are executable, and thus produce verifiable accounts of how neural

processes give rise to behaviour via the emergence of representations and decision-making strategies. In other words, neural models can serve as a computational theory of language use (Baroni, 2022). They can produce data-driven algorithmic explanations of the principles that guide linguistic interaction.

Evidence that neural language models can be used as effective tools for the language sciences is increasingly strong, yet the space of possible applications is still vastly uncharted. This thesis explores novel ways of using neural language models as models of human language use, with the goal of establishing new methods and enabling new research directions for a wide variety of language scientists, from historical linguists, sociolinguists, and lexicographers to neuroscientists and psycholinguists. Insights from this line of research further inform and stimulate the development of language processing technologies that more faithfully reproduce human linguistic behaviour.

1.3 Main contributions and overview

This first chapter has so far presented the perspective and motivation that have informed this thesis. **Chapter 2** develops this introduction by tying it to the general background, which includes a review of arguments in favour of using artificial neural networks as models of language use as well as an introduction to language modelling and a brief history of language model architectures.

The rest of the thesis is then structured into three parts: word usage (Part 1), utterance comprehension (Part 2), and utterance production (Part 3). These are the three main aspects of language use that the studies in this thesis address. Part 1 and 2 are connected by a focus on language comprehension; Part 2 and 3 are about entire utterances rather than individual words. What ties all three parts together are (i) a computational approach to the study of language use based on neural language models, (ii) a pragmatic perspective on the role of context in shaping linguistic communication, and (iii) the scientific goal of learning about quantitative aspects of linguistic behaviour, while generating insights about the state of the art of computational modelling.

Part 1: Word Usage

Part 1 explores ways of using neural language models to study word usage and interpretation as modulated by sentential context. I present two novel ways of using neural language models to obtain lexical representations as a function of a word's context of occurrence. The two proposed types of lexical representation are evaluated in terms of their suitability for semantic change analysis, an established computational linguistic task which requires capturing word meaning with its nuanced context-determined modulations. Overall, this part of the thesis delivers useful tools to linguists and social scientists, while shedding light on language

models' ability to interpret words—an ability which underpins general machine reading comprehension.

Chapter 3 introduces the relevant background, which includes word meaning representations in NLP, computational approaches to the modelling of word meaning change over time, and a review of the word 'definition modelling' task.

Chapter 4 presents a new methodological approach which consists of extracting, grouping, and analysing contextualised word representations from neural language models. This is the first unsupervised approach to lexical semantic change that makes use of contextualised neural word representations. I propose several metrics to quantify a word's degree of semantic change with this type of lexical representation, create a new evaluation dataset of human similarity judgements, and use them to show that contextualised representations and their detected semantic shifts correlate with human intuitions. The proposed approach captures synchronic phenomena such as word senses and syntactic functions, literal and metaphorical word usage, as well as diachronic linguistic processes related to the narrowing and broadening of word meaning over time. This chapter also contains an extensive evaluation of the proposed approach on four indo-european languages. I test contextualised representations obtained using different neural architectures (LSTM and Transformer language models), training corpora, and change detection algorithms. Based on empirical findings, I make practical recommendations for the deployment of contextualised word representations as detectors of word meaning change.

Chapter 5 presents a second novel approach to semantic change analysis that relies on human-readable word definitions generated by neural language models. I show that word definitions automatically generated with a specialised language model can serve as interpretable representations for polysemous words. The generated definitions are evaluated via human judgements of definition quality and by comparison against human word usage similarity judgements. Generated definitions are in most cases accurate, understandable, and approximate human judgements of word usage similarity better than the previously introduced contextualised word representations. I also demonstrate how word sense definitions obtained through a simple modification of this main approach can be used to produce interpretable descriptions of diachronic relations between word senses, thus providing explanations for meaning changes observed in diachronic text corpora.

Part 2: Utterance Comprehension

In Part 2, I use neural language models to study aspects of utterance comprehension as a function of the relevant discourse context. In particular, I obtain estimates of contextualised *surprisal* (or information content) from neural language models and use these to test psycholinguistic theories of utterance production which postulate speakers' monitoring of information rate due to its effect on comprehension effort. Findings from these analyses challenge established

hypotheses of rational use of the communication channel, especially in dialogic settings. Overall, however, they confirm that strategies of utterance production can be described as efficiently containing surprisal and thus the comprehension effort of interlocutors. Faithful modelling of information transmission strategies in humans, to which this part of the thesis contributes, informs the development of improved technologies for natural language understanding and generation.

Chapter 6 begins with an introduction of the relevant background: linguistic theories of audience-awareness and collaborative (joint production and comprehension) effort, the role of surprisal in psycholinguistic accounts of language processing, and the information-theoretic notions that underlie surprisal theory. Then, it presents a method to obtain estimates of utterance surprisal using autoregressive neural language models. While previous work estimates surprisal without taking discourse context into account, the proposed approach allows measuring utterance surprisal both in and out of context, thereby making it possible to quantitatively inspect the role of context in reducing predictive uncertainty over next utterances. This chapter defines the main information theoretic measures of surprisal, describes the computational models that produce the surprisal estimates, and evaluates these estimates intrinsically, in terms of their ability to fulfil expectations about context-sensitive language processing.

Chapter 7 applies the proposed surprisal estimation method to study patterns of information transmission in English texts as well as spoken and written dialogues. Central tenets of the classic information-theoretic model of communication, which predict rational strategies of information transmission within a noisy channel, are put to test. In particular, I revisit the entropy rate constancy and the uniform information density hypotheses. While the results of this analysis complement and support prior evidence of rational production strategies in texts, they also suggest that the noisy-channel model of communication, coupled with rationality assumptions, may paint too simplistic a picture for dialogue, where two (or more) speakers have to monitor and coordinate information transmission strategies on the fly. The chapter continues by focusing on entropy rate constancy and uniform information density in task-oriented dialogue, with a focus on how task-determined contextual units affect patterns of information transmission. I identify theoretically motivated contextual units over which participants may deploy strategies of information management and compression, and observe that dialogue participants' production strategies are more accurately described as rational when analysed within topically coherent and reference-specific contextual units rather than within entire dialogues.

Chapter 8 moves the focus to open-ended dialogue, for which the weakest empirical evidence of rational communication strategies was observed in the previous chapter: in purely conversational settings, speakers seem to progressively reduce their collaborative effort over time. I test the hypothesis that speakers use construction repetition (i.e., the repeated use of particular configurations of structures and lexemes) as a strategy for surprisal mitigation—in particular, by

padding the more information dense parts of their utterances with progressively less information dense lexical bundles. I define a new information-theoretic measure that captures the *facilitating effect* of constructions on utterance processing in dialogue. The findings of this analysis confirm that constructions exhibit lower surprisal than other expressions and that their surprisal decreases with repetition—leading to an overall decrease in information rate over the course of a dialogue.

Part 3: Utterance Production

The goal of Part 3 is to lay the foundations for using neural language generators as models of utterance production. While natural language generation systems are widely deployed in real-world applications, evidence that they faithfully reproduce aspects of human linguistic behaviour is very scarce. I start by analysing variability in human production, a characterising aspect of language production that is often overlooked in natural language generation research. I propose a statistical framework to quantify variability and to assess language generators' alignment to the production variability observed in humans. Just like, in Part 2, neural models of comprehension allowed studying audience-aware production strategies, here I use language generators as models of production to study human comprehension behaviour. I define novel measures of utterance predictability that are complementary with the probabilistic surprisal measures used in Part 2 and test their psychometric predictive power, showing that they can predict and explain human acceptability judgements and reading times. I conclude Part 3 by collecting insights from the rest of the thesis into a formal framework for artificial simulations of human-like—efficient and communicatively effective—language production behaviour. Overall, beyond its importance to linguistic and psycholinguistic research, this part of the thesis complements Part 2 in informing and stimulating the development of language processing technologies that more faithfully reproduce human linguistic behaviour.

Chapter 9 introduces the relevant background beyond what is already presented in the general background and the background of Part 2. This includes natural language generation—with a focus on automatic evaluation—and covers the relation between surprisal and predictability in expectation-based theories of language processing.

Chapter 10 presents a statistical framework to quantify variability in language production. Using datasets that collect multiple human utterances given the same production context, I measure human variability across four production tasks, providing empirical evidence for qualitative expectations about the open-endedness of different communicative scenarios. I then assess neural text generators' compliance to the levels of variability observed in human data and find that, overall, generators are well calibrated—which suggests that they can begin to be used to study aspects of language production in humans. The chap-

ter also draws strong connections between variability and uncertainty, resulting in tools to measure sequence-level uncertainty in language models and to evaluate the statistical fit of natural language generation systems.

Chapter 11 builds on the finding that neural language generators reproduce human production variability. I use the technical tools and the framework introduced in the previous chapter to design novel measures of utterance predictability. The proposed measures are based on the idea that comprehenders form expectations and reason over alternative utterances that speakers may have but did not produce in a given communicative context. I use neural text generators to obtain sets of plausible alternatives and measure utterance predictability in terms of an utterance's distance from the alternative set. I assess the robustness of the proposed measure of predictability, '*information value*', to different configurations of the underlying estimators, and then I demonstrate its ability to predict human comprehension behaviour in the form of acceptability judgements and reading times. Not only does information value possess the favourable property of being inherently sequence-level and interpretable—unlike aggregates of token-level surprisal; it also has stronger psychometric predictive power for acceptability judgements in spoken and written dialogue and is complementary to surprisal as a predictor of reading times.

Chapter 12 collects insights from the rest of the thesis into a conceptual framework for efficient and communicatively effective—i.e., pragmatic—natural language generation in variably complex communicative scenarios. The framework relies on four main notions: context, communicative goals, production and comprehension costs, and communicative utility. I define these notions formally and, in two case studies, I provide suggestions for their operationalisation in classic generation tasks. I argue that human-like, pragmatic linguistic behaviour emerges as a result of reasoning about context, goals, costs, and utility, and I discuss possible promising directions towards pragmatic natural language generation systems that learn to reason about these concepts.

The thesis ends with an overall summary and a brief discussion of the implications of my main contributions (**Chapter 13**).

1.3.1 List of publications

This thesis is largely based on ideas, methods, and findings presented in the following nine papers.

Part 1: Word Usage

1. Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computa-*

tional Linguistics, pages 3960–3973, Online. Association for Computational Linguistics.

2. Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

3. Mario Giulianelli, Iris Luden, Raquel Fernández, and Andrey Kutuzov. 2023. Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Part 2: Utterance Comprehension

4. Mario Giulianelli and Raquel Fernández. 2021. Analysing Human Strategies of Information Transmission as a Function of Discourse Context. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 647–660, Online. Association for Computational Linguistics.

5. Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is Information Density Uniform in Task-Oriented Dialogues?. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

6. Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2022. Construction Repetition Reduces Information Rate in Dialogue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), pages 665–682, Online only. Association for Computational Linguistics.

Part 3: Utterance Production

7. Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What Comes Next? Evaluating Uncertainty in Neural Text Generators Against Human Production Variability. To appear in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Republic of Singapore. Association for Computational Linguistics.

8. Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. Information Value: Measuring Utterance Predictability as Distance from Plausible Alternatives. To appear in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Republic of Singapore. Association for Computational Linguistics.

9. Mario Giulianelli. 2022. Towards Pragmatic Production Strategies for Natural Language Generation Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7978–7984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chapter 2

General background

In this chapter, I provide an overview of the main concepts that underlie all the studies presented in the thesis. The chapter is split into two sections. The first section introduces *artificial neural networks* and highlights why they are suitable as models of language use. The second section describes *language models*, the key technology that powers the methods and analyses in this thesis.

2.1 Artificial neural networks as models of language use

This thesis consists of a series of studies conducted with the aim of contributing new supporting evidence to the claim that *artificial neural networks* are a class of models that is suitable to predict and explain human linguistic behaviour, and that they can therefore be used as a linguistic formalism.

Artificial neural networks belong to the tradition of Parallel Distributed Processing (PDP; McClelland et al., 1986). PDP approaches can provide a mechanistic explanation of the principles underlying human language use as they model cognitive processes as arising from simple interactions of neurons through synaptic connections. Indeed, since their inception, they have been used to model a wide range of cognitive tasks, such as past-tense learning (Rumelhart and McClelland, 1986), the Stroop effect (Cohen et al., 1990), and serial recall (Botvinick and Plaut, 2006). Furthermore, the PDP paradigm sits well with the language-as-action tradition introduced in the previous chapter: knowledge that governs processing is stored in the strengths of the network connections and is acquired gradually through experience.

Now, the toolkit of the language and cognitive sciences already includes good explanatory models. Cognitive process models, for example, which typically consist of simple and interpretable components, can explicitly encode human cognitive priors resulting from (biological) neural architecture, as well as environmental

statistics, development, and evolutionary pressure (Ma and Peters, 2020). So why should we use artificial neural networks instead? Cognitive process models require hand-crafted features and decision rules, which (i) introduce potentially unverifiable assumptions and (ii) may prevent scaling to complex and high-dimensional tasks. In contrast, deep neural networks—the modern connectionist models—can solve highly complex tasks with very few assumptions and little-to-no feature engineering (recurrent neural networks and Transformers, e.g., perform almost as well as humans in a variety of language understanding tasks), they can serve as models of *task learning* (Rogers and McClelland, 2004; Rumelhart et al., 1986; Lake et al., 2017; Kruschke, 1992), and they have the potential to explain how neural processes give rise to behaviour, via the *emergence* of representations and decision strategies.

Deep learning models have indeed been shown to be compatible with modern linguistic frameworks such as the Rational Speech Act, a social cognition approach to language use (Frank and Goodman, 2012b; Goodman and Frank, 2016) which is compatible with the story of goal-directed action and perception presented in the previous chapter, and which has found notable empirical support as a computational level description of language use—for instance, in experiments on the interpretation of referring expression, figurative language, vagueness and embedded implicatures. Combinations of neural models and RSA have been employed to study the role of both listeners and speakers (Monroe et al., 2017; Newman et al., 2020; Andreas and Klein, 2016) in linguistic interaction—allowing the RSA framework to generalise to more complex communicative scenarios. Neural networks can provide an utterance space for cognitive agents to reason over; they can construct contextual representations and, with respect to these, verify utterances; and, especially when trained with reinforcement learning curricula, they can reason about goals, learn utility and cost functions, and use them to perform action planning.

2.2 Language modelling

In this section, I will introduce language models, the key technology used throughout this series of studies. First, I will present language modelling as a 'task', or general problem formulation, and then I will survey the main types of language models as of today, focusing on their varying ability to make use of contextual information—an aspect that is crucial to their deployment in this thesis. There exist countless resources that describe language models at virtually any level of detail; this is intended to be a targeted, non-exhaustive, and—with a few necessary exceptions—informal introduction.

It is not uncommon for a person to find themselves in a conversation and suddenly realise that they can accurately predict what their interlocutor is going to say next, even without them completing their utterance.

(1) We should really visit...

If they know their interlocutor well and have a shared history of relevant interactions, they may be able to guess that the utterance continues as follows:

(2) We should really visit Inverness and see the Loch Ness monster!

The two speakers have previously talked about this topic and it is shared knowledge that Inverness may be a fun place to visit over the weekend, so hearing the word 'visit' restricts the space of possible sentence continuations to 'Inverness', 'Loch Ness', and perhaps a few other plausible alternatives such as 'your parents' or 'the art gallery'. It is not necessary to know the interlocutor to predict that the sentence will not continue with a verb—for example, 'study'—as that would make the sentence ungrammatical, nor with a noun phrase such as 'the desk' because that would make the sentence semantically implausible. There can be exceptions: unlikely or apparently implausible continuations may also occur. For example, the speaker may decide to say:

(3) We should really visit the monster this weekend!

Because 'the monster' is a surprising yet meaningful continuation, this utterance has a certain probability to carry a comedic effect—at least between the two interactants.

Predicting which word is likely to come next given a situational and interactional context is one of the central problems in natural language processing and computational linguistics, and it is referred to as *language modelling*. Being able to form expectations (e.g., by assigning a probability) to next words has many practical applications. In their Introduction to Natural Language Processing, Computational Linguistics, and Speech Processing, Dan Jurafsky and James Martin mention a few (2009). For example, having a model of likely sequences of words can help speech recognition systems identify words in noisy and ambiguous input, it can help grammatical error correction systems detect and correct mispellings, and it is essential for good translation systems.

Models that assign probabilities to sequences of words are called **language models** or, in short, **LMs**. Typically, these consist of an autoregressive factorisation of the probability of sequences, with conditional nextword probabilities

predicted by a statistical model θ:

$$\begin{aligned} \boldsymbol{p}_\theta(\text{We should really visit the monster}) = \\ \boldsymbol{p}_\theta(\text{monster}|\text{We should really visit the}) \times \\ \boldsymbol{p}_\theta(\text{the}|\text{We should really visit}) \times \\ \ldots \times \\ \boldsymbol{p}_\theta(\text{We}|\text{<s>}) \end{aligned} \tag{2.1}$$

where '<s>' is a special symbol indicating the beginning of a sentence. We will refer to the input of a language model as the *context*. The context is a random variable X that can take a specific value $X = x$. A context x can be a sequence of words such as that in Equation 2.1, it can be a source sentence in machine translation, a sequence of dialogue turns, or any natural language string input to the language model. Given a context, the language model predicts a distribution over the next sequence, a random variable Y. This, in turn, can take a specific value $Y = y$ such as a translated sentence, a dialogue turn, or any word sequence that plausibly follows the context x. Using this terminology and notation, language modelling can be defined as the task of predicting a probability distribution over the countably infinite set of natural language strings given, as a context, another natural language string:

$$p(Y|X = x) \ . \tag{2.2}$$

Given the string 'We should really visit', a good language model will assign a higher probability to more likely grammatical continuation 'the' than to 'study'. Moreover, a language model that has been exposed to the history of interactions between the interlocutors will not find the word 'Inverness' surprising as a continuation. Perhaps it will assign higher probability to it than to 'Belfast' even though 'Belfast' is a more frequent word. Generally speaking, the most common way of obtaining language models that make accurate next word predictions is by optimising the underlying statistical model via maximum likelihood estimation. Given a text corpus, the model is presented with one sentence at a time and its parameters are adapted in such a way that the probability of the observed word sequences is maximised. There are some drawbacks to this approach—the fact that *Belfast* is not observed in the data does not mean it is not a plausible continuation—and they will be discussed in Part 3 of this thesis. Nevertheless, most language models these days are optimised through this simple learning objective. The exact way the parameters θ are updated depends on the type of statistical model. In the next section, I will review five types of model: n-gram, (feed-forward) neural network, RNN, LSTM, and Transformer language models. The focus will be on how contextualisation is achieved in each of them, highlighting how improvements in a model's ability to capture contextual cues are strongly related to the overall quality of the model.

2.2.1 A brief history of language models

We have defined language modelling as the task of predicting the conditional probability distribution $P(Y|X = x)$ over word sequences. I have also mentioned that this complex task is typically re-expressed as the easier, more tractable problem of predicting the probability of a sequence $y = (w_1, \dots w_m)$ one word at a time (Equation 2.1). More precisely, the joint probability $P(Y = w_1 \dots w_m|X = x)$—which can be spelled out as $P(Y_1 = w_1 \cap Y_2 = w_2 \cap \dots \cap Y_m = w_m|X = x)$—is re-expressed, through the chain rule, as the multiplication of conditional next word probabilities, for each word in the sequence:

$$\begin{aligned} &p(Y = w_1 \dots w_m|X = x) = \\ &p(Y_1 = w_1|X = x) \times \\ &p(Y_2 = w_2|X = x\ w_1) \times \\ &\dots \times \\ &p(Y_m = w_m|X = x\ w_1 \dots w_{m-1}) \end{aligned} \tag{2.3}$$

One naive approach to determining this probability involves using relative frequency counts. Given a corpus of texts, we can define the probability of w_i given $X = (x\ w_1\ \dots\ w_{i-1})$ as the number of occurrences of the word sequence $(x\ w_1\ \dots\ w_{i-1})$ followed by the word w_i divided by total number of occurrences of $(x\ w_1\ \dots\ w_{i-1})$:

$$p(Y_i = w_i|X = x\ w_1 \dots w_{i-1}) = \frac{\text{count}(x\ w_1 \dots w_{i-1}\ w_i)}{\text{count}(x\ w_1 \dots w_{i-1})} \tag{2.4}$$

The problem with this naive approach is that the contextual sequence instantiating the random variable X can be a long word sequence, and long word sequences are rare by nature. Language is recursive and creative, so there exist infinite sequences one can form with words from a natural language vocabulary. Moreover, language is in constant change and new sequences are created all the time. Consequently, any particular context may have never existed before, especially if it consist of a long sequence.

A simple alternative approach is to assume that the probability of the next word only depends of the last n words in the contextual sequence. This is commonly referred to as an n-th order Markov assumption. The probability $p(Y_i = w_i|X = x\ w_1 \dots w_{i-1})$ can be then approximated as $p(Y_i = w_i|X = w_{i-n} \dots w_{i-1})$. The subsequence $(w_{i-n} \dots w_{i-1})$ is typically referred to as an n-gram. The advantage of this approach is that, when n is a low number such as 2, 3, or 4, many n-grams are no longer rare and their counts can be estimated from a corpus. Language models that approximate the conditional probability of the next word as its probability given the previous n words are called ***n*-gram language models**. Since some n-grams will still inevitably be rare (e.g., the frequency of the bigram 'AI alignment' in a text corpus collected in 2010 is probably

close to zero), so-called *smoothing* techniques have been developed throughout the years to redistribute the probability of seen n-grams to unseen ones (e.g., Jelinek and Mercer, 1980; Katz, 1987; Kneser and Ney, 1995; Gale and Sampson, 1995), thus preventing that a probability of zero is assigned to sequences containing unseen n-grams.

With our without smoothing, n-gram language models have important flaws. First, while the Markov assumption makes it possible to compute relative counts and approximate sequence probabilities, estimating likelihoods by only looking at the previous few words inevitably leads to inaccurate next word predictions. Topical and logical coherence, for example, cannot be guaranteed as they require looking further backwards in the context; even more local phenomena related to grammaticality and selectional preferences cannot be properly modelled as they might often require tracking long-distance dependencies between words (e.g., it is impossible to assign the right probability to the base form of an English verb vs. its third person form if the subject of that verb occurs more than n words away). A second important drawback has to do with storing n-gram counts for large text corpora. The amount of memory required simply scales with the number of texts observed, making n-gram language models particularly memory-inefficient. A third weakness is that the modelling of word and n-gram usage is performed independently for each word in the vocabulary and for every possible n-gram. There is no re-sharing of information between word and n-gram features. For example, because they have no notion of semantic similarity between words and phrases, n-gram language models have no way of re-using the relative counts of the phrase 'lexical semantic change' when the phrase 'lexical semantic shift', which has been never observed in the training corpus, appears at test time.

Neural network language models tackle the last two issues, resulting in more efficient and generalisable architectures. The key idea is to use high-dimensional vectors to represent words, rather than their simple identity (i.e., a symbol). Given an n-gram context, the feature vector of the corresponding n words can be concatenated and used as an abstract representation for the context. In this way, neural LMs can exploit the semantic similarity between 'change' and *shift* to bootstrap a similar context representation for 'lexical semantic change' and 'lexical semantic shift'. Relative counts are no longer needed. They are implicitly captured by a series of (layers of) affine transformations, which re-express context representations in alternative high-dimensional spaces, from which next probabilities can be calculated. The following is a simple neural network language model with a context window of size 4:

$$\boldsymbol{c}_i = \text{concat}(\boldsymbol{e}_{w_{i-3}}, \boldsymbol{e}_{w_{i-2}}, \boldsymbol{e}_{w_{i-1}}, \boldsymbol{e}_{w_i}) \tag{2.5}$$

$$\boldsymbol{h}_i = \tanh(W^{(h)}\boldsymbol{c}_i + \boldsymbol{b}^{(h)}) \tag{2.6}$$

$$\boldsymbol{p}_{i+1} = \text{softmax}(W^{(p)}\boldsymbol{h}_i + \boldsymbol{b}^{(p)}) \tag{2.7}$$

where $\boldsymbol{e}_{w_i}$ (a numerical vector representing word w_i), $W^{(h)}$, $W^{(p)}$, $\boldsymbol{b}^{(h)}$, and $\boldsymbol{b}^{(p)}$

are trainable model parameters. Maximum likelihood estimation with neural networks is achieved through error back-propagation. After the network has processed the context vector and predicted from it next word probabilities, its parameters (both the word representations and the transformation matrices) are updated with a simple learning rule: the parameters are changed in such a way that the next word w_i^* observed in the training corpus becomes more likely:

$$\mathscr{L} = -\log p_{i+1}(Y_{i+1} = w_{i+1}^* | X = x) \tag{2.8}$$

We will encounter this quantity again in Part 2, where we will refer to it as *information content* or *surprisal* (Shannon, 1948) and interpret it as quantifying the degree of unexpectedness of a linguistic signal in context. Surprisal is a mathematical operationalisation of signal predictability which has been empirically successful for the study of human language processing—and this is not a coincidence: artificial neural networks were originally developed as model of the cognitive mechanisms underlying human behaviour.

It is not a coincidence either that the next type of language model we will discuss relies on an artificial neural architecture developed by a cognitive scientist, John Elman, with the goal of giving representation to time (and thus to sequences of observations) in neural networks (Elman, 1990). Elman networks are neural networks equipped with *context units*, artificial memory units designed to store information from prior stages of processing. These units enable the network to access and reuse previously stored information, enhancing its ability to process sequential data effectively. The resulting neural networks are commonly referred to as Recurrent Neural Networks, or RNNs. **RNN language models** introduce one crucial modification with respect to the neural network LMs in Equations 2.5 to 2.7. The hidden unit $\boldsymbol{h}_i$ (Equation 2.6) is computed as a function of the current word representation as well as the representation of the computations involved in the processing of the sequence up until the current word, as stored in the context-memory unit $\boldsymbol{h}_{i-1}$:

$$\boldsymbol{h}_i = \tanh(W^{(x)}\boldsymbol{e}_{w_i} + W^{(h)}\boldsymbol{h}_{i-1} + \boldsymbol{b}^{(x)}) \tag{2.9}$$

$$\boldsymbol{p}_{i+1} = \text{softmax}(W^{(p)}\boldsymbol{h}_i + \boldsymbol{b}^{(p)}) \tag{2.10}$$

Crucially, the previous context unit $\boldsymbol{h}_{i-1}$ does not only represent the previous word w_{i-1}, nor the previous n words, but rather all the words in the contextual sequence up to the current word w_i. This is the fundamental difference between RNN language models and simple feedforward neural network language models (and, of course, n-gram language models), which, at least in principle, allows RNNs to condition next word predictions on contexts of unlimited length.

In practice, however, RNNs are only able to model short contexts. To learn dependencies between distant words, back-propagation must travel back in time, through $W^{(h)}$, until the beginning of the word sequence. At each time step, the magnitude of the gradient decreases and it eventually vanishes before it reaches

the furthest portions of the context. To obviate this *vanishing gradient problem*, two other cognitive scientists, Sepp Hochreiter and Jürgen Schmidhuber, developped the Long Short-Term Memory, or LSTM (Hochreiter and Schmidhuber, 1997). **LSTM language models** include an explicit memory cell which controls the gradient by design and prevents it from vanishing after a few time steps.

$$\boldsymbol{u}_i = \tanh(W^{(xu)}\boldsymbol{e}_i + W^{(hu)}\boldsymbol{h}_{i-1} + \boldsymbol{b}^{(u)} \tag{2.11}$$

$$\boldsymbol{i}_i = \sigma(W^{(xi)}\boldsymbol{e}_i + W^{(hi)}\boldsymbol{h}_{i-1} + \boldsymbol{b}^{(i)}) \tag{2.12}$$

$$\boldsymbol{f}_i = \sigma(W^{(xf)}\boldsymbol{e}_i + W^{(hf)}\boldsymbol{h}_{i-1} + \boldsymbol{b}^{(f)}) \tag{2.13}$$

$$\boldsymbol{o}_i = \sigma(W^{(xo)}\boldsymbol{e}_i + W^{(ho)}\boldsymbol{h}_{i-1} + \boldsymbol{b}^{(o)}) \tag{2.14}$$

$$\boldsymbol{c}_i = \boldsymbol{i}_i \odot \boldsymbol{u}_i + \boldsymbol{f}_i \odot \boldsymbol{c}_{i-1} \tag{2.15}$$

$$\boldsymbol{h}_i = \boldsymbol{o}_i \odot \tanh(\boldsymbol{c}_i) \tag{2.16}$$

$$\boldsymbol{p}_{i+1} = \text{softmax}(W^{(p)}\boldsymbol{h}_i + \boldsymbol{b}^{(p)}) \tag{2.17}$$

The input, output, and forget *gates* ($\boldsymbol{i}_i$, $\boldsymbol{i}_o$, and $\boldsymbol{i}_f$) have an intuitive interpretation: they regulate the amount of contextual information that is remembered and forgotten by the network, by modulating the output of hidden and recurrent computations. Addressing the vanishing gradient problem allows LSTM language models to learn long-distance dependencies between words in the context. As a result, the next word predictions of LSTM models can track topical and logical coherence, as well as syntactic structures and selectional preferences with a wide scope over surface forms.

While LSTM language models are able to take into account large contexts, they still compress the representation of the entire context into a single recurrent unit, thus limiting the amount of information that can be propagated through distant portions of the context. Instead of making use of a single recurrent unit, so-called neural *attention mechanisms* keep track of the representations of each word in the context and combine them when predicting the current word.[1] One of the most popular forms of attention is the *scaled dot-product attention* (Vaswani et al., 2017). Given an observation (w_1 w_2 … w_m) and the word w_i at a given position i within the sequence, the amount of attention on any other word w_j in the sequence is determined by the query vector $\boldsymbol{q}_i$ and the key vector $\boldsymbol{k}_j$. The output representation for word w_i is then computed by weighting the value vector $\boldsymbol{v}_j$ of each word w_j (including w_i itself) by the attention weight a_{ij}:

$$a_{ij} = \text{softmax}(\boldsymbol{q}_i \boldsymbol{k}_j^\top) \tag{2.18}$$

$$\boldsymbol{h}_i = \sum_{j=1\ldots m} a_{ij}\boldsymbol{v}_j \tag{2.19}$$

[1] These are more precisely referred to as *self-attention* mechanisms as they condition computations on other parts of the same observation. Attention is a more general mechanism in a neural network that allows the model to learn to make predictions by attending to a given set of data.

Query, key, and value vectors are learned parameters. This makes attention-based neural models able to jointly optimise the selective attention weights (through $\boldsymbol{q}$ and $\boldsymbol{k}$) and the representation of each input input (through value vectors $\boldsymbol{v}$). Transformer neural models (Vaswani et al., 2017) build on this fundamental mechanism by adding two main modifications. First, rather than only computing the attention once, Transformers use a *multi-head attention* mechanism which computes the scaled dot-product attention multiple times in parallel and combines them through concatenation and linear transformation. Second, Transformers consist of multiple layers of multi-head attention, interleaved with normalisation and feed-forward neural layers. The resulting neural models are the backbone of **Transformer language models**.

Similarly to LSTM models, Transformer LMs typically include static word embeddings as input, or first-layer representations. Furthermore, because Transformers have no in-built notion of time (attention is a function over a set of context positions), the input layer typically also includes *positional embeddings* to allow the model to learn different attention patterns according to the position of a word in the context. Transformer LMs are currently the most commonly used neural architecture for language models, and they take two main forms. *Encoders* have as a goal to learn a representation of the input sequence (mainly by learning to predict the probability of the current word given its left and right context: $\boldsymbol{p}_i = \text{softmax}(W\boldsymbol{h_i} + \boldsymbol{b})$; this is called *masked language modelling* because the model tries to predict words for a set of masked positions in the sequence). *Decoders*, on the other hand, learn to predict the next word, in a more classic language modelling setup: $\boldsymbol{p}_{i+1} = \text{softmax}(W\boldsymbol{h_i} + \boldsymbol{b})$; this is called *autoregressive* or *causal language modelling*. Encoders and decoders can be combined, such as in the original Transformer model (Vaswani et al., 2017), or used separately: BERT is perhaps the most prominent example of an encoder (Devlin et al., 2019b), while GPT-2 is certainly the most important example of a decoder-only language model (Brown et al., 2020b).

A last property of the Transformer language models which will be used throughout this thesis is that they are *pre-trained* on massive amounts of texts, rather than trained on specific corpora of interest, and that they possess a very large number of parameters. While these may seem only quantitative differences between these and previous models (with a few exceptions in the recurrent family (e.g., Peters et al., 2017, 2018; McCann et al., 2017), the scale-up effects of model size and data quantity on the LM behaviour can be easily appreciated at a qualitative level. Large pre-trained Transformer are by far the most accurate language models of language use ever developed.

Part One

Word Usage

There are many ways of conceptualising and describing the meaning of words. A common way to describe the meaning of a word is using a natural language definition such as those that can be found in dictionaries. If one looks up the word 'alignment' in the Cambridge Dictionary, for example, the first entry will read *'an arrangement in which two or more things are positioned in a straight line or parallel to each other'*. This is a helpful and human-friendly way of describing the meaning of 'alignment': one can read it and immediately get a sense of how the word can be *used*. For most entries, dictionaries also include usage examples to demonstrate in which linguistic contexts a word might appear—for example, 'the problem is happening because the wheels are out of alignment with each other'. This is not the only meaning of 'alignment'. The Cambridge Dictionary also includes a second *sense* of the word: *'an agreement between a group of countries, political parties, or people who want to work together because of shared interests or aims'*—along with the following usage example: 'New alignments are being formed within the business community'. The lexicographers who redacted the current version of the Cambridge Dictionary have decided that these are the two main senses of the word, the two main ways in which the word is used by English speakers. The first is a literal interpretation of 'alignment', which describes the relative position of objects in physical space; the second can be considered a metaphorical extension of the first word sense, which describes the relative position of entities in any type of abstract space.

A useful alternative way to think of the word 'alignment'—and, in fact, of any word in natural language—is as a form of behaviour speakers use to transmit specific bits of information encoded in their brain. Said differently, a word is a set of instructions speakers use to allow their audience to reconstruct the information they wanted to transmit, thus re-creating the intended meaning in their brain (Traugott, 2017). If a speaker has information regarding the symmetry in the relative position of some objects in physical space, or of some entities in abstract space, they can use the word 'alignment' to ensure their audience can reconstruct, and thus becomes aware of this information. Provided that they share the same basic conventions, the audience will make use of the current situation, the environment of the linguistic interaction, and the nature of the 'aligned' entities to follow the instructions and reconstruct the intended bits of information. Under this lens, the static sets of word senses and word sense definitions that we are used to finding in dictionaries offer limited descriptions of word meaning. If a word is a set of instructions to create meaning and the execution of the instructions varies depending on their context of occurrence, the same word can take a myriad of—wildly or subtly—different meanings. This indeed happens when the word is produced by different communities of speakers, whose communicative environments and language use differ from each other; but it also occurs for different individual speakers within the same community as well as for the same speaker across communicative contexts.

Communicative needs are always contextual and always renewing, and as

context changes, speakers tend to parsimoniously make use of the repertoire of instructions they have mastered. Among linguists, for example, the term 'alignment' can be used to refer to the system of morphosyntactic rules that distinguishes between the arguments of transitive verbs and those of intransitive verbs, as well as to describe the mechanisms by which speakers adapt to each other's posture, speech rate, or word usage (a phenomenon also often termed *entrainment*). In recent years, computational linguists and artificial intelligence researchers have had to learn yet another meaning of the word, emerged within a certain American community of AI enthusiasts. They use 'alignment' to describe the (still imperfect) correspondence between the goals and interests of humans and those of AI systems. This last case really does exemplify the arbitrary nature of the relation between signifier and signified de Saussure and others have explored (de Saussure, 1972). What it also demonstrates is that communicative needs change across linguistic communities: over space—both geographical and socio-cultural—and over time. How, then, can we describe word meaning in a way that accounts for polysemy (the property of having multiple senses at the same time), for nuanced context modulation within the same overarching word sense, and without relying on ad-hoc decisions made by restricted groups of experts? Each usage of the word 'alignment' should be described differently depending on its context of occurrence. Some usages can be more similar to each other, and other totally different. Good *word representations* allow comparing word usages, judging their similarity, and—when necessary—grouping usages that are similar enough to form a coherent group.

The first part of this thesis explores ways of using neural language models to represent word meaning. In Chapter 4, I present an approach to extract contextualised word representations from neural language models. The proposed method produces sub-symbolic representations: these are essentially vectors of continuous values that encode word meaning in an abstract high-dimensional space. Contextualised neural model representations fulfil the desiderata described above but have one main disadvantage: they are not directly interpretable by humans. In Chapter 5, I address this limitation by using neural language models to generate word definitions from usage examples. These are context-dependent representations that can be compared and grouped on the basis of their similarity, while being directly understandable by humans. Unlike static dictionary definitions, these are determined by unique contexts of word occurrence, and the resulting groups (word senses) are obtained in a data-driven way. I will use diachronic word meaning change as a case study for these two types of lexical representations as tracking word meaning over time requires capturing nuanced context-determined modulations and is thus a suitable test-bed for the proposed approaches.

Chapter 3

Background

3.1 Word representations in NLP

Understanding of the meaning of individual words underpins general machine reading comprehension. For this reason, developing algorithms to capture lexical meaning has been a long-standing goal of artificial intelligence and natural language processing research. As discussed in the introduction to Part 1, to be ecologically plausible and useful in the face of semantic variation and change (see also Section 3.2), lexical representations should be context-sensitive and data-driven.

Early approaches to representing word meaning include large-scale databases like WordNet (Miller, 1995), which represent words and their underlying concepts in terms of their relations to other concepts. This taxonomical approach has many advantages—for example, it allows to easily retrieve synonyms, antonyms, hyponyms, and hypernyms of a word of interest—but it offers a static, encyclopedic view over the meaning of words that does not allow capturing their constantly evolving nature. A new, more flexible paradigm for lexical representations emerged when NLP researchers started to design statistical operationalisations of the **distributional hypothesis** (Harris, 1954; Firth, 1957). The distributional hypothesis states that semantically similar words possess similar linguistic distributions—i.e., they occur in similar contexts. Distributional approaches exploit statistical regularities in the contexts of word occurrences to induce semantic representations, and they are the most common nowadays. In the following sections, I will present the three main types of distributional word representations: form-based, sense-based, and usage-based representations.

3.1.1 Form-based word representations

Among the first successful methods for the statistical modelling of the distribution of words are **count-based methods** (Baroni and Lenci, 2010; Turney and Pantel,

2010). Given the vocabulary V of a language, these methods count the number of times that each word in V occurs in the vicinity of every other word in a corpus of texts (for example, within a surrounding window of ten tokens). These counts can be stored in ordered lists, or *word vectors*, such that every word in the vocabulary is assigned its own vector and that every element in the vector corresponds to another word in the vocabulary. Why does this result in good word representations? As an example, the vectors constructed by looking at the occurrences of the words 'alignment' and 'arrangement', or 'alignment' and 'calibration', will be more similar to each other than the vectors for 'alignment' and 'cat' or 'alignment' and 'theatre'.

Concurrent and slightly later approaches exploit the same intuition but instead of counting co-occurrences, they rely on **predictive models**—most often, neural networks—to learn which words are likely to appear in the same sentential context (Collobert and Weston, 2008; Turian et al., 2010; Collobert et al., 2011; Mikolov et al., 2013; Pennington et al., 2014). The weights of the neural networks corresponding to each word in the vocabulary can be then extracted and used directly as word representations. The resulting word vectors exhibit the same core properties of count-based vectors, in that representations of semantically similar words are closer in high-dimensional space than those of semantically dissimilar ones. Predictive models have been for years the most popular method for the modelling of word meaning since they yield word representations that possess three essential characteristics (Boleda, 2020): (i) they are learned without supervision from unprocessed natural language data, (ii) their multidimensionality encompasses various subtle aspects of meaning, although not necessarily easy to interpret (Boleda and Erk, 2015), and (iii) their continuous nature allows capturing graded semantic phenomena like word similarity, synonymy, lexical priming, and selectional preferences.

3.1.2 Sense-based word representations

An obvious issue with form-based approaches is that they conflate all meanings of a word into a single static representation (Camacho-Collados and Pilehvar, 2018). For example, while occurrences of the word 'bank' may have different syntactic and semantic properties depending on the word interpretation intended in a given context (e.g., 'bank' as a building, 'bank' as an institution, or 'to bank'), the different types of context corresponding to varying word interpretations are not distinguished for the creation of form-based representations. Therefore, all occurrences of the word 'bank', regardless of their meaning in context, will be assigned the same word representation. While, in practice, this issue is particularly problematic for highly polysemous words such as 'bank' or 'play' and less so for monosemous ones (there are very few examples of these, 'monosemous' being a good one), it affects the overall validity of form-based representations: as discussed in the introduction to Part 1, the meaning of virtually all words

(including, e.g., monosemous words and even many function words) is modulated by context.

Luckily, that *word forms* (or *lemmas*, such as 'bank') should be taken as the unit of meaning in distributional appraoches does not follow from the distributional hypothesis. **Sense-based representations** aim to capture the different usages of a word (for example, the different meanings of polysemous words) by separately modelling the contexts in which each usage type, or *word sense*, is likely to occur. Usages of the word 'bank' as a financial institution and as a building will be more similar to one another than to uses of 'bank' that refer to the land alongside a river or a lake. But how can different word senses, i.e., the most frequent or most prototypical usages of a word, be distinguished from one another? One common way is through word sense disambiguation (WSD) algorithms (McCarthy, 2009; Navigli, 2009). Supervised approaches to WSD use labelled training data to train statistical models that predict the correct sense of a word in new contexts. Unsupervised solutions typically measure similarities between contexts without the need of training data. Pioneering work by Schütze (1998), for example, exploits the concept of *second-order co-occurrence*: two contexts of a target word are assigned to the same sense if the words the two context co-occur with, in turn, are similar. Once the contexts of a word are grouped into different senses, a separate distributional representation can be learned for each sense.

The inherent issue with sense-based representations is that determining when two usages of the same word are similar enough to belong to the same sense is an arbitrary decision (Cruse, 1995; Kilgarriff, 1997; Kintsch, 2007), especially in lack of a consensus on how word meaning is represented in the human brain (Klein and Murphy, 2001; Falkum and Benito, 2015). Two sense-specific representations can be similar in some dimensions (e.g., for 'bank', those related to money or finance) and different in others (e.g., the abstract vs. physical axis of 'bank')—and the notion of similarity should be taken as graded. (Boleda, 2020). Moreover, judging whether two usages of a word are similar is largely and ultimately a subjective matter (Brown, 2008; Erk et al., 2013).

Overall, while they still build on problematic assumptions about the nature of word meaning unsupervised word sense disambiguation approaches provide technical solutions which are more aligned to the desiderata for lexical representations presented in the introduction to Part 1, as compared with form-based representations.

3.1.3 Usage-based word representations

We have so far discussed approaches that take word forms and word senses as meaning units, and we have seen that both these units of meaning are at odds with what we know about how humans use and interpret words. If a word is a set of instructions to create meaning, within a specific communicative and situational

context, each individual act of word production and interpretation should be represented with its own unique signature. To some extent, unsupervised word sense disambiguation approaches do give importance to individual usages and contexts of occurrence. The previous section describes an example method where individual contexts of word usage are compared and grouped based on their pairwise similarity (Schütze, 1998).

Usage-based word representations completely bypass word senses and consider individual word occurrences as a meaning unit. Early approaches are grounded in compositional distributional semantics, in that they compute usage-specific word representations by composing vectors for word forms with vectors representing the context of occurrence (Erk and Padó, 2008, 2010). These representations are explicitly modulated by the context in which a word appears, allowing for more nuanced and flexible representations. In recent years, a new strand of usage-based word representations emerged together with advances in neural language models. These representations, often also referred to as *contextualised representations* or *contextualised embeddings*, are essentially the neural activations of a language model that result from processing a target word within its context of occurrence.

Contextualised word representations have been shown to encode lexical meaning accurately and dynamically: they are good predictors of usage similarity judgements (Pilehvar and Camacho-Collados, 2019) and, without explicit supervision, they perform on a par with state-of-the-art word sense disambiguation models in disambiguation tasks (Wiedemann et al., 2019). In Chapters 4 and 5, I will present two types of usage-based word representations designed to capture a context-sensitive and evolving notion of word meaning. I will evaluate them against human judgements similarity of usage similarity and demonstrate their applicability to the analysis of diachronic word usage and meaning change.

3.2 Diachronic word usage change

Naming a concrete or abstract entity with a given word is a convention established among a linguistic community, a behaviour which is recognised by most members as an encoding of the same entity. Over time, speakers change the way they refer to concepts and objects in the world. Communities are complex dynamic systems in constant evolution—speakers interact, move to new places, they grow older, die, and are born. Moreover, individual speakers are non-deterministic decision-makers who may use, more or less deliberately, different types of behaviour to achieve the same goals, and who continuously learn new ways of fulfilling old intents. Concepts and objects change, too, over time—for example, as a result of advancements in technology, science, and culture—and to be successfully communicated, they demand new forms of behaviours. In sum, language is a living and dynamic system which is constantly evolving, and word usage change is an

inherent part of this evolution. Words and phrases that were commonly used in the past may have different meanings today, and new words and phrases are continually being created to communicate new concepts and ideas.

This section will focus on diachronic **lexical semantic change**—i.e., the change of word meaning over time.

3.2.1 Computational modelling of word usage change

Computational modelling of word usage change can be approached with diverse scientific motivations. This section discusses the most important axes along which computational approaches to lexical semantic change vary.

New conventions may arise around an existing word, and as a result, the word can start to refer to completely unrelated concepts (e.g., the meaning of 'coach' as an instructor emerged in the 18th century and is a homonym of the much older meaning of 'coach' as a horse-drawn carriage). Often, the new word meaning is related to an existing one (as it is the case for 'coach' as a railway carriage vs. a single-decker bus). Since the late 19th century, lexical semantic change has been a subject of research in diachronic linguistics. Blank (1997) provides an extensive historical overview; here, I briefly describe two popular taxonomies of **semantic change types** to provide an intuition of the possible diachronic trajectories the meaning of a word can take. Hermann Paul's taxonomy is the first example. It classifies semantic change into three primary types: meaning *specialisation* (or *narrowing*), *generalisation* (or *broadening*), and *transfer*——the latter encompassing what would later be referred to as *metaphorisation* and *metonymisation* (Paul, 1886). Leonard Bloomfield enumerates nine change types, including *metaphor*, *metonymy*, *synecdoche*, *hyperbole*, and *meiosis*, as well as *pejoration* and *amelioration* (Bloomfield, 1933). Sense-based and usage-based approaches have the potential to discern between change types, yet the development of novel methods that detect changes of varying nature relies heavily on the availability of annotated datasets that accurately record instances of change and their corresponding change types (Sander, 2023).

Moreover, meaning change trajectories unravel at different timescales. Changes can manifest within a few weeks or months, and they can be transient. For instance, in January 2020, journalists typically employed the word 'virus' to refer to the general concept of a replicating infectious agent. Soon after, the usage of 'virus' underwent a rapid and profound process of specialisation, eventually denoting a new more specific referent, SARS-CoV-2 (Montariol et al., 2021). Such instances of lexical semantic change are often referred to as cases of *short-term meaning change*. In contrast, there exist more enduring and less abrupt instances of semantic change. For example, the usage of the word 'mouse' as a pointing device for personal computers gradually gained prominence throughout the latter half of the twentieth century (Wijaya and Yeniterzi, 2011). Computational models can adopt positions along a continuum of **temporal granularity**, ranging

from days and weeks to decades and centuries.

Computational approaches can, furthermore, seek to provide varying **levels of description** and explanation, relating lexical semantic change to different types of determining factors. Sociolinguists may focus on analysing the sociocultural factors influencing word usage, such as whether speakers employ 'coach' to refer to a bus or a carriage. Psycholinguists may be more interested in cognitive factors that lead speakers to assign the same word form to unrelated or distantly related meanings, such as using 'coach' to denote a sports instructor or trainer. Others may be interested in social pressures and study the relationship between community structure and semantic change (Noble et al., 2021), or in explaining change as a result of cognitive pressures (Grewal and Xu, 2021). These approaches offer complementary yet independent accounts of semantic change. Integrated accounts combining social and cognitive explanations are still relatively scarce in the literature.

Another classically important distinction is between **semasiological** and **onomasiological** approaches to lexical semantic change (Geeraerts, 1997). Semasiological modelling explores the changes in meaning associated with a particular word over time (e.g., Gulordava and Baroni, 2011; Hamilton et al., 2016). For example, researchers may study how the meaning of the term 'alignment' has transformed from its earliest occurrence in the field of NLP to the present day. Onomasiological approaches, on the other hand, centre around changes in the collection of words associated with a particular concept across different time periods (e.g., Betti and Van den Berg, 2014; Sommerauer and Fokkens, 2019). This type of modelling can detect cases of lexical replacements and named entity change (Szymanski, 2017) as well as, for example, that *'the problem of engineering AI systems with goals and values similar to humans' goals and values'* has only recently started to be referred to as 'alignment'.

Lastly, there tend to be methodological differences between studies focusing on **semantic change discovery** and studies focusing on **semantic change detection** (or tracking). The aim of semantic change discovery is to identify across the entire vocabulary (or a restricted subset, such as all nouns) words that have undergone meaning change over a specific period (Kurtyigit et al., 2021). In contrast, semantic change detection examines the meaning trajectories of curated lists of target words with the goal of determining whether and to what extent the meaning of those words has changed over time (Kutuzov et al., 2018; Tahmasebi et al., 2018). While this axis should be rather thought of as a continuum, the technical distinction between the two types of studies is more pronounced. Modern approaches often rely on computationally intensive language models which may not be suitable for large-scale semantic change discovery. Therefore, there is growing interest in methods that combine the accuracy of modern LM-based approaches with the efficiency necessary for tracking the entire vocabulary (Zamora-Reina et al., 2022; Tahmasebi et al., 2022).

3.3 Definition modelling

The task of generating human-readable word definitions, as found in dictionaries, is commonly referred to as *definition modelling* or *definition generation* (for a review, see Gardner et al., 2022). This task will become relevant in Chapter 5, where I will present a method to generate contextualised word definitions and to use them as human-readable word representations.

The original motivation for the definition modelling task has been the interpretation, analysis, and evaluation of otherwise obtained abstract word representation spaces. Definition generation systems, however, also have practical applications in lexicography, language acquisition, and sociolinguistics, as well as within NLP (Bevilacqua et al., 2020). The task was initially formulated as the generation of a natural language definition given an embedding—a single distributional representation—of the target word, or *definiendum* (Noraset et al., 2017). The first formulation of definition modelling was soon replaced by the task of generating a contextually appropriate word definition given a target word embedding and an example usage (Gadetsky et al., 2018; Mickus et al., 2022)—a task description which takes into account that word meaning is always contextually determined.

Generating definitions from vector representations is not the most natural formulation of definition modelling. Ni and Wang (2017) and Mickus et al. (2019) treat the task as a sequence-to-sequence problem: given an input sequence with a highlighted word, generate a contextually appropriate definition. In Chapter 5, we will follow this approach.

3.3.1 Methods

Methods that address this last formulation of the task are typically based on a pre-trained language model deployed on the definienda of interest in a natural language generation (NLG) setup (Bevilacqua et al., 2020). Generated definitions can be further improved by regulating their degree of specificity via specialised LM modules (Huang et al., 2021), by adjusting their level of complexity using contrastive learning training objectives (August et al., 2022), or by supplementing them with definitional sentences extracted directly from a domain-specific corpus (Huang et al., 2022). In Chapter 5, we will compare the results obtained with our proposed approach to the state-of-the-art specificity-tuned definition generator proposed by Huang et al. (2021).

3.3.2 Evaluation

Generated definitions are typically evaluated with standard NLG metrics such as BLEU, NIST, ROUGE-L, METEOR, or MoverScore (e.g., Huang et al., 2021; Mickus et al., 2022), using precision@k on a definition retrieval task (Bevilacqua

et al., 2020), or measuring semantic similarity between sentence embeddings obtained for the reference and the generated definition (Kong et al., 2022). Because reference-based methods are inherently flawed (for a discussion, see Mickus et al., 2022), qualitative evaluation is almost always presented in combination with these quantitative metrics. In Chapter 5, we will evaluate generated definitions with automatic metrics and by collecting human judgements.

Chapter 4

Contextualised neural word representations

The content of this chapter is based on the following publications:

> Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.
>
> Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

For the first study, the three authors jointly produced the research idea. Mario performed the experiments with Marco and Raquel's supervision. Mario wrote the article, Marco and Raquel provided guidance and contributed to the writing. For the second study, the two authors jointly produced the idea for the article, which is an extension of the first study. Mario and Andrey performed the experiments and wrote the article. The text in this chapter overlaps with that of the original publications.

4.1 Introduction

In the fourteenth century, the words 'boy' and 'girl' referred, respectively, to a male servant and a young person of either sex (Oxford English Dictionary). By the fifteenth century, a narrower usage had emerged for 'girl', designating exclusively female individuals, whereas by the sixteenth century 'boy' had lost its servile connotation and was more broadly used to refer to any male child, becoming the masculine counterpart of 'girl' (Bybee, 2015). Word meaning is indeed in constant mutation and, since correct understanding of the meaning of individual words underpins general machine reading comprehension, it has become increasingly relevant for computational linguists to detect and characterise lexical semantic change with the aid of quantitative and reproducible procedures—e.g., in the form of laws of semantic change (Dubossarsky et al., 2015; Xu and Kemp, 2015; Hamilton et al., 2016).

Most recent studies have focused on *lexical semantic change detection* (or *shift detection*), the task of deciding whether and to what extent the concept evoked by a word has changed between time periods (e.g., Gulordava and Baroni, 2011; Kim et al., 2014; Kulkarni et al., 2015; Del Tredici et al., 2019; Hamilton et al., 2016; Bamler and Mandt, 2017; Rosenfeld and Erk, 2018). This line of work relies mainly on distributional semantic models which produce one abstract representation for every word form (see Section 3.1 in the previous background chapter). However, aggregating all senses of a word into a single representation is particularly problematic for semantic change modelling as word meaning hardly ever shifts directly from one sense to another, but rather typically goes through polysemous stages (Hopper et al., 1991). This limitation has motivated recent work on word sense induction across time periods (Lau et al., 2012; Cook et al., 2014; Mitra et al., 2014; Frermann and Lapata, 2016; Rudolph and Blei, 2018; Hu et al., 2019). Word senses, however, have shortcomings themselves as they are a discretisation of word meaning, which is continuous in nature and modulated by context to convey ad-hoc interpretations (Brugman, 1988; Kilgarriff, 1997; Paradis, 2011).

In this study, we propose a usage-based approach to lexical semantic change modelling, where sentential context modulates lexical meaning "on the fly" (Ludlow, 2014). We present a novel method that (i) exploits a pre-trained neural language model to obtain contextualised representations for every occurrence of a word of interest, (ii) clusters these representations into *usage types*, and (iii) measures change along time. More concretely, we make the following contributions:

- We present the first unsupervised approach to lexical semantic change that makes use of state-of-the-art contextualised word representations.
- We propose several metrics to measure semantic change with this type of representation. Our code is available at `https://github.com/glnmario/cwr4lsc`.

- We create a new evaluation dataset of human similarity judgements on more than 3K word usage pairs across different time periods, available at `https://doi.org/10.5281/zenodo.3773250`.
- We show that both the model representations and the detected semantic shifts are positively correlated with human intuitions.
- Through in-depth qualitative analysis, we show that the proposed approach captures synchronic phenomena such as word senses and syntactic functions, literal and metaphorical word usage, as well as diachronic linguistic processes related to the narrowing and broadening of word meaning over time.

Overall, our study demonstrates the potential of using contextualised word representations for modelling and analysing lexical semantic change and opens the door to further work in this direction.

In an extension of this study, presented in Section 4.6, we evaluate this approach more extensively, across four languages, by participating in the SemEval-2020 Shared Task 1 on lexical semantic change detection.

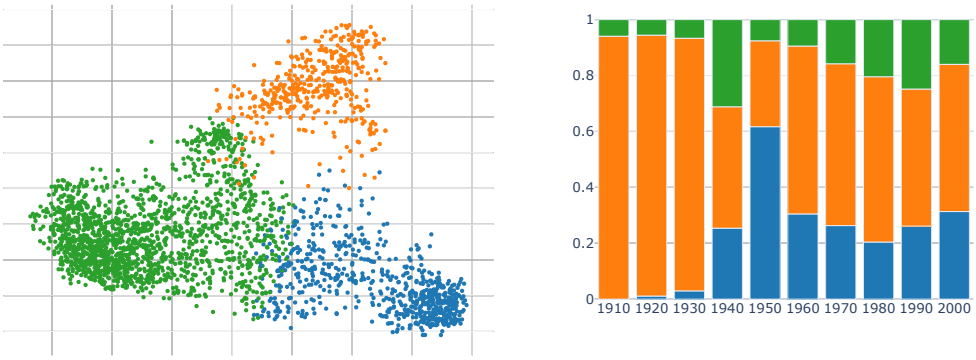
4.2 A usage-based approach to lexical semantic change modelling

We introduce a usage-based approach to lexical semantic change modelling which relies on contextualised representations of unique word occurrences (*usage representations*). First, given a diachronic corpus and a list of words of interest, we use a pre-trained language model (BERT; Devlin et al., 2019a) to compute usage representations for each occurrence of these words. Then, we cluster all the usage representations collected for a given word into an automatically determined number of partitions (*usage types*) and organise them along the temporal axis. Finally, we propose three metrics to quantify the degree of change undergone by a word.

4.2.1 Language model

We produce usage representations using the BERT language model (Devlin et al., 2019a), a multi-layer bidirectional Transformer encoder trained on masked token prediction and next sentence prediction on the BooksCorpus (800M words) (Zhu et al., 2015) and on English text passages extracted from Wikipedia (2,500M words). We use the smaller *base-uncased* version of BERT, with 12 layers, 768 hidden dimensions, and 110M parameters.[1]

[1]We rely on Hugging Face's implementation of BERT (available at `https://huggingface.co/bert-base-uncased`).



(a) PCA visualisation of the usage representations.

(b) Probability-based usage type distributions over time.

Figure 4.1: Usage representations and usage type distributions generated based on occurrences of the word *atom* in COHA (Davies, 2012). Colours indicate inferred usage types.

4.2.2 Usage representations

Given a word of interest w and a context of occurrence $s = (v_1, ..., v_i, ..., v_n)$ with $w = v_i$, we extract the activations of all of BERT's hidden layers for sentence position i and sum them dimension-wise. Alternative layer aggregation strategies are presented and evaluated in Section 4.6.

The set of N usage representations extracted for the occurrences of w in a given corpus can be expressed as the usage matrix $\mathbf{U}_w = (\mathbf{w}_1, \ldots, \mathbf{w}_N)$. For each usage representation in the usage matrix $\mathbf{U}_w$, we store the context of occurrence (a 128-token window around the target word) as well as a temporal label $\mathbf{t}_w$ indicating the time interval of the usage.

4.2.3 Usage types

Once we have obtained a word-specific matrix of usage vectors $\mathbf{U}_w$, we standardise it and cluster its entries using k-Means.[2] This step partitions usage representations into clusters of similar usages of the same word, or *usage types* (see Figure 4.1a). It is thus directly related to automatic word sense disambiguation (Schütze, 1998; Pantel and Lin, 2002; Manandhar et al., 2010; Navigli and Vannella, 2013, among others).

For each word independently, we automatically select the number of clusters K_w that maximises the silhouette score (Rousseeuw, 1987), a metric of cluster quality which favours intra-cluster coherence and penalises inter-cluster similarity, without the need for gold labels. For each value of K, we execute 10 iterations of Expectation Maximization to alleviate the influence of different initialisation

[2]Other clustering methods are also possible. For this first study, we choose the widely used k-Means (and rely on the implementation available through the *scikit-learn* library).

values (Arthur and Vassilvitskii, 2007). The final clustering for a given K is the one that yields the minimal *distortion* value across the 10 runs, i.e., the minimal sum of squared distances of each data point from its closest centroid. We experiment with $K \in [2, 10]$ and choose this range heuristically: we forgo $K = 1$, as k-Means and the silhouette score are ill-defined for this case, while keeping the number of possible clusters manageable computationally. This excludes the possibility that a word has a single usage type. Alternatively, we could use a measure of intra-cluster dispersion for $K = 1$, and consider a word monosemous if its dispersion value is below a threshold d (if the dispersion is higher than d, we would discard $K = 1$ and use the silhouette score to find the best $K \geq 2$). There also exist clustering methods that select the optimal K automatically, such as DBSCAN or Affinity Propagation (as tested, e.g., by Martinc et al., 2020). They nevertheless require method-specific parameter choices which indirectly determine the number of clusters.

By counting the number of occurrences of each usage type k in a given time interval t (we refer to this count as $\text{freq}(k, t)$), we obtain frequency distributions $\mathbf{f}_w^t$ for each interval under scrutiny:

$$\mathbf{f}_w^t \in \mathbb{N}^{K_w} : \mathbf{f}_w^t[k] = \text{freq}(k, t) \quad k \in [1, K_w] \tag{4.1}$$

When normalised, frequency distributions can be interpreted as probability distributions over usage types $\mathbf{u}_w^t : \mathbf{u}_w^t[k] = \frac{1}{N_t}\mathbf{f}_w^t[k]$. Figure 4.1b illustrates the result of this process.

4.2.4 Quantifying semantic change

We propose three metrics for the automatic quantification of lexical semantic change using contextualised word representations. The first two (*entropy difference* and *Jensen-Shannon divergence*) are known metrics for comparing probability distributions. In our approach, we apply them to measure variations in the relative prominence of coexisting usage types. We conjecture that these kinds of metric can help detect semantic change processes that lead to broadening or narrowing (i.e., to increase or decrease, respectively, in the number or relative distribution of usage types). The third metric (*average pairwise distance*) only requires a usage matrix $\mathbf{U}_w$ and the temporal labels $\mathbf{t}_w$ (Section 4.2.2). Since it does not rely on usage type distributions, it is not sensitive to possible errors stemming from the clustering process.

Entropy difference (ED). We propose measuring the uncertainty (e.g., due to polysemy) in the interpretation of a word w in interval t using the normalised entropy of its usage distribution $\mathbf{u}_w^t$:

$$\eta(\mathbf{u}_w^t) = \log_{K_w}\left(\prod_{k=1}^{K_w} \mathbf{u}_w^t[k]^{-\mathbf{u}_w^t[k]}\right) \tag{4.2}$$

To quantify how uncertainty over possible interpretations varies across time intervals, we compute the difference in entropy between the two usage type distributions in these intervals:

$$\mathrm{ED}(\mathbf{u}_w^t, \mathbf{u}_w^{t'}) = \eta(\mathbf{u}_w^{t'}) - \eta(\mathbf{u}_w^t) \ . \tag{4.3}$$

We expect high ED values to signal the broadening of a word's interpretation and negative values to indicate narrowing.

Jensen-Shannon divergence (JSD). The second metric takes into account not only variations in the size of usage type clusters but also *which clusters* have grown or shrunk. It is the Jensen-Shannon divergence (Lin, 1991) between usage type distributions:

$$\mathrm{JSD}(\mathbf{u}_w^t, \mathbf{u}_w^{t'}) = \mathrm{H}\left(\frac{1}{2}\left(\mathbf{u}_w^t + \mathbf{u}_w^{t'}\right)\right) - \frac{1}{2}\left(\mathrm{H}\left(\mathbf{u}_w^t\right) - \mathrm{H}\left(\mathbf{u}_w^{t'}\right)\right) \tag{4.4}$$

where H is the Boltzmann-Gibbs-Shannon entropy. Very dissimilar usage distributions yield high JSD whereas low JSD values indicate that the proportions of usage types barely change across periods.

Average pairwise distance (APD). While the previous two metrics rely on usage type distributions, it is also possible to quantify change by bypassing the clustering step into usage types. We do so by calculating the average pairwise distance between usage representations in different periods t and t':

$$\mathrm{APD}(\mathbf{U}_w^t, \mathbf{U}_w^{t'}) = \frac{1}{N^t \cdot N^{t'}} \sum_{\mathbf{x}_i \in \mathbf{U}_w^t,\ \mathbf{x}_j \in \mathbf{U}_w^{t'}} d(\mathbf{x}_i, \mathbf{x}_j) \tag{4.5}$$

where $\mathbf{U}_w^t$ is a usage matrix constructed with occurrences of w only in interval t. We experiment with cosine, Euclidean, and Canberra distance.

Generalisation to multiple time intervals. The presented metrics quantify semantic change across pairs of time intervals (t, t'). When more than two intervals are available, we measure change across all contiguous intervals ($m(\mathbf{U}_w^t, \mathbf{U}_w^{t+1})$, where m is one of the metrics), and calculate the mean and maximum metric values.[3] The *mean* is indicative of semantic change across the entire period under consideration, while the *max* pinpoints the pair of successive intervals where the strongest shift has occurred.

[3]The Jensen-Shannon divergence can also be measured with respect to $T > 2$ probability distributions (Ré and Azad, 2014): $\mathrm{JSD}\left(\mathbf{u}_w^1, \ldots, \mathbf{u}_w^T\right) = \mathrm{H}\left(\frac{1}{T}\sum_{i=1}^T \mathbf{u}_w^i\right) - \frac{1}{T}\sum_{i=1}^T \mathrm{H}\left(\mathbf{u}_w^i\right)$. However, this definition of the JSD is insensitive to the order of the temporal intervals and yields lower correlation with human semantic change ratings (cf. Section 4.4.2) than the pairwise metrics.

4.3 Data

We examine word usages in a large diachronic corpus of English, the Corpus of Historical American English (COHA, Davies, 2012). COHA covers two centuries (1810–2009) of language use and includes a variety of genres, from fiction to newspapers and popular magazines, among others. In this study, we focus on texts written between 1910 and 2009, for which a minimum of 21M words per decade are available, and discard previous decades, where texts are less balanced per decade.

As target words, we use the 100 lemmas annotated with semantic shift scores by Gulordava and Baroni (2011). These shift scores are human judgements collected by asking five annotators to quantify the degree of semantic change undertaken by each word—shown out of context—from the 1960's to the 1990's. We exclude *extracellular* as it only appears in three decades of COHA; all other words appear in at least 8 decades, with a minimum and maximum frequency of 191 and 108,796 respectively. We refer to the resulting set of 99 words and corresponding shift scores as the 'GEMS dataset' or the 'GEMS words', as appropriate.

We collect a contextualised representation for each occurrence of these words in the second century of COHA using BERT as described in Section 4.2.2. This results in a large set of usage representations, ∼1.3M in total, which we cluster into usage types using k-Means and silhouette coefficients (Section 4.2.3). We use these usage representations and usage types in the evaluation and the analyses presented in Sections 4.4 and 4.5.

4.4 Correlation with human judgements

Before using our proposed method to analyse lexical semantic change, we assess how its key components compare with human intuition. We test whether the clustering into usage types reflects human similarity judgements (Section 4.4.1) and to what extent the degree of change computed with our metrics correlates with shift scores provided by humans (Section 4.4.2).

4.4.1 Evaluation of usage types

The clustering of contextualised representations into usage types is one of the main steps in our method (see Section 4.2.3). It relies on the similarity values between pairs of usage representations created by the language model. To quantitatively evaluate the quality of these similarity values (and thus, by extension, the quality of usage representations and usage types), we compare them to human raters' similarity judgements.

New dataset of similarity judgements. We create a new evaluation dataset, following the annotation approach of Erk et al. (2009, 2013) for rating pairs of usages of the same word. Since we need to collect human judgements for pairs of usages, annotating the entire GEMS dataset would be extremely costly and time consuming. Therefore, we limit the scope of the annotation to a subset of words. For each shift score value s in the GEMS dataset, we sample a word uniformly at random from the words annotated with s. This results in 16 words. To ensure that our selection of usages is sufficiently varied, for each of these words, we sample five usages from each of their usage types (the number of usage types is word-specific) along different time intervals—one usage per 20-year period over the century.[4] All possible pairwise combinations are generated for each target word, resulting in a total of 3,285 usage pairs.

We use the crowdsourcing platform Figure Eight[5] to collect five similarity judgements for each of these usage pairs. To control the quality of the similarity judgements, we select Figure Eight workers from the pool of most experienced contributors, we require them to be native English speakers and to have completed a test quiz consisting of 10 similarity judgements.[6] The compensation scheme for the raters is based on an average wage of 10 USD per hour. Annotators are shown pairs of usages of the same word: each usage shows the target word in its sentence, together with the previous and the following sentences (67 tokens on average). Annotators are asked to assign a similarity score on a 4-point scale, ranging from *unrelated* to *identical*, as defined by Brown (2008) and used, e.g., by Schlechtweg et al. (2018). Figures 4.2 and 4.3 show the full instructions given to the annotators and Figure 4.4 illustrates a single annotation item. A total of 380 annotators participated in the task. The inter-rater agreement, measured as the average pairwise Spearman's correlation between common annotation subsets, is 0.59. This is in line with previous annotation efforts such as those by Schlechtweg et al. (2018), who report agreement scores between 0.57 and 0.68.

Results. To obtain a single human similarity judgement per usage pair, we average the scores given by five annotators. We encode all averaged human similarity judgements for a given word in a square matrix. We then compute similarity scores over pairs of usage vectors output by BERT[7] to obtain analogous matrices per word and measure Spearman's rank correlation between the human- and the machine-generated matrices using the Mantel test (Mantel, 1967).

We observe a significant ($p < 0.05$) positive correlation for 10 out of 16 words,

[4]When a usage type does not occur in a time interval, we uniformly sample an interval from those that do contain occurrences of that usage type.

[5]`https://www.figure-eight.com`, recently acquired by Appen (`https://appen.com`).

[6]For this purpose, I manually annotated 170 usage pairs.

[7]For this evaluation, BERT is given the same variable-size context as the human annotators. Vector similarity values are computed as the inverse of Euclidean distance, because k-Means relies on this metric for cluster assignments.

with correlation coefficients ranging from 0.13 to 0.45. Table 4.1 presents the correlation coefficients and p-values obtained for each word. Overall, this is an encouraging result, which indicates that BERT's word representations and similarity scores (as well as our clustering methods which build on them) correlate to a substantial extent with human similarity judgements, and thus provides a promising empirical basis for our approach.

	ρ	p
federal	0.131	0.001
spine	0.195	0.032
optical	0.227	0.003
compact	0.229	0.002
signal	0.233	0.008
leaf	0.252	0.001
net	0.361	0.001
coach	0.433	0.007
sphere	0.446	0.002
mirror	0.454	0.027
card	0.358	0.055
virus	0.271	0.159
disk	0.183	0.211
brick	0.203	0.263
virtual	-0.085	0.561
energy	0.002	0.990

Table 4.1: Spearman's correlation results per target word: BERT vs. human similarity judgements.

4.4.2 Evaluation of semantic change scores

We now quantitatively assess the semantic change scores yielded by the metrics described in Section 4.2.4 when applied to BERT usage representations and to the usage types created with our approach. We do so by comparing our scores to the human shift scores in the GEMS dataset. For consistency with this dataset, which quantifies change from the 1960's to the 1990's as explained in Section 4.3, we only consider these four decades when calculating our scores. Using each of the metrics on representations from these time intervals, we assign a semantic change score to all the GEMS words. We then compute Spearman's rank correlation between the automatically generated change scores and the gold standard shift values.

Overview

Each question includes two sentences. Both sentences contain a **target word** between double brackets, as in: [[target]]. Your task is to rank the similarity of the two usages of the target word according to the following scale:

1. unrelated
2. distantly related
3. closely related
4. identical

IMPORTANT: your task is to evaluate the similarity of the two usages of the same word, **not the similarity of the two sentences** in general.

If you are unable to choose a label because you do not understand the sentences, select the option "cannot decide". Please try to use this option as little as possible!

An example

You will see two sentences. Both contain the target word marked by double brackets; in this example it's the word [[current]].

Read the sentences carefully:

- in any case , it 's not a question of electrocution . we can arrange a relay which will break the [[current]] at the instant of application of weight . if the robot should place his weight on it , he wo n't see .
- already , while it was still a blueprint , they were proud of their idea , of its simple clean lines and undeniable originality -- it owed nothing in its conception to any of the [[current]] models of revolutionary strategy . the japanese red army comrades , whadi haddad and his pflp contingent , even the matchless " carlos " could only admire .

And then select how similar the two usages of the word [[current]] are:

1. unrelated
2. distantly related
3. closely related
4. identical
5. (cannot decide)

You can choose only **one** label. Please try to use the option "cannot decide" as little as possible.

Why do texts look weird?

The sentences you'll read don't look like they were taken from a book. This is because they have gone through some text processing. You should not be concerned nor influenced in your decisions by the fact that:

- all words are lowercase (written in small letters), even proper names or the pronoun "I"
- whitespaces may appear where you don't expect them (e.g. before a comma) and may sometimes not appear where you'd expect them (e.g. between words)
- strange characters and words occasionally appear
- some words are misspelled
- a few words are missing
- the target word may appear multiple times (*but your judgement should be about the occurrence signalled by the*[[ ]] *marker*)

Please simply **ignore** these aspects while labelling!

Figure 4.2: Annotation instructions for usage similarity judgements (part 1).

What do the similarity labels mean? More examples

Let's now look at examples for all four labels. Remember that you are evaluating the similarity of two word usages—not the overall similarity of the two sentences!

1. How similar are these two usages of [[current]]?

- prices of the leading issues . considering past earnings records , are apparently on a conservative basis measured by [[current]] market valuations in other groups . on the other hand there is no particular speculative incentive for operations in this group , with all signs pointing to a lower volume of sales in the last half of the year .
- one of the weirdest was the disappearance of anchovies off the coast of peru . why this happened is still unclear . one theory is that the cause was the 1972 - 73 invasion of a warm-water [[current]] called el nino , which upset the ecology of the coldwater humboldt current , drastically reducing the supply of 119 economics in plain english plankton and other nutrients on which anchovies (as well as whales) feed .

UNRELATED.
In the first sentence, "current" means being most recent or occurring at the present time. In the second sentence, "current" refers to a flow of water within a lake or an ocean. These two meanings have no properties in common; it is not possible to explain one usage in terms of the other.

2. How similar are these two usages of [[current]]?

- it is quite possible to arrive at the right conclusions for the wrong reasons , just as it is possible to ignore history but not to repeat it . thus , the summers book represents an important [[current]] of thought in the u.s. military , which rightly argues that the vietnam defeat was not the fault of the military ; never again should young americans be sent into battle without public backing and a clear definition of the goals of the military engagement .
- get busy with that l-tube ! If you do n't have it apart , cleaned , and together again before the day is out , i 'll coagulate your brains with alternating [[current]] . " not a robot moved !

DISTANTLY RELATED.
In the first sentence, "current" refers to a feeling or idea that exists within a group of people. In the second sentence, "current" refers to a flow of electricity. Although these two meanings of "current" are different, they are related as they do share some properties: for example, both currents of thought and electronic current can *flow*, and they are both often the result of an interplay of forces. This is why connecting the two meanings in a sentence results in a perhaps sophisticated but understandable statement: (A) "a new current of thought is flowing through the circuits of parliament".

3. How similar are these two usages of [[current]]?

- one of the weirdest was the disappearance of anchovies off the coast of peru . why this happened is still unclear . one theory is that the cause was the 1972 - 73 invasion of a warm-water [[current]] called el nino , which upset the ecology of the coldwater humboldt current , drastically reducing the supply of 119 economics in plain english plankton and other nutrients on which anchovies (as well as whales) feed
- he said he had survived by managing in a stupor to drag himself into a windowless shed behind his house. col. michael wiener , an israeli army doctor , said many of the survivors in the valley below the lake , such as in souboum , may simply have been in an air [[current]] that did not have any poison , while someone standing only a few yards away may have been killed. dr. weiner , the head of a 17-member rescue unit that came to cameroon team to nkamba , about 100 miles northeast of bamenda .

CLOSELY RELATED.
In the first sentence, "current" refers to a flow of water within a lake or an ocean. In the second sentence, "current" refers to a steady flowing movement of air. These two meanings are closely related: both refer to a steady and continuous flowing movement of some physical element. As in the previously encountered example (A), we can construct a sentence that relates the two usages: (B) "I can't remember whether El Nino is the name of an ocean or a wind current". Note, however, that the meanings are still ultimately different: a flow of water is not a flow of air.

4. How similar are these two usages of [[current]]?

- dell 's shares , on the other hand , go for 26 times projected 2004 earnings-but its business is three times as profitable as apple 's . The company 's supporters say [[current]] profits matter little because jobs has proved time and time again that he can create new products and trailblaze markets . that may be so , but as transamerica portfolio manager chris bonavico , who does n't own apple 's stock , notes , " Apple will remain a company that is neat from a product and consumer standpoint but crap from an investor standpoint .
- prices of the leading issues . considering past earnings records , are apparently on a conservative basis measured by [[current]] market valuations in other groups . on the other hand there is no particular speculative incentive for operations in this group , with all signs pointing to a lower volume of sales in the last half of the year .

IDENTICAL.
In both sentences, "current" means being most recent, up-to-date, or occurring at the present time. The two meanings are identical because they share virtually all properties, as can be seen from the following example. In the sentence (C) "I can't remember whether Donald Trump is the current president or vice-president of the United States", the meanings of "current" are ultimately the same regardless of whether Trump is president or vice-president. (Note the difference with respect to the constructed sentence (B) above.)

Now you're ready to start!

Figure 4.3: Annotation instructions for usage similarity judgements (part 2).

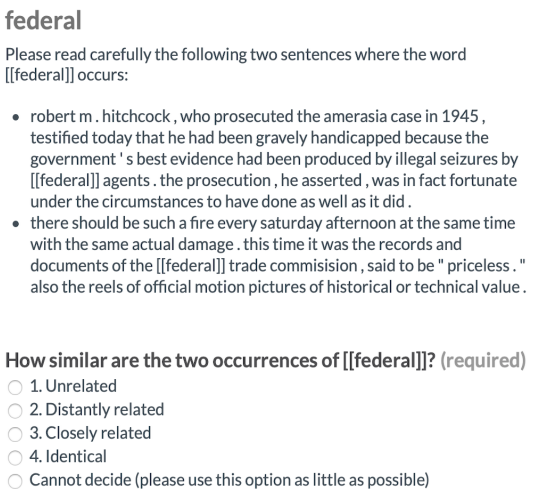


Figure 4.4: An annotation item, as it appears on the Figure Eight crowdsourcing platform.

Results. Table 4.2 shows the Spearman's correlation coefficients obtained using our metrics, together with a frequency baseline (the difference between the normalised frequency of a word in the 1960's and in the 1990's). The three proposed metrics yield significant positive correlations. This is again a very encouraging result regarding the potential of contextualised word representations for capturing lexical semantic change.

As a reference, we report the correlation coefficients with respect to GEMS shift scores documented by the authors of two alternative approaches: the count-based model by Gulordava and Baroni (2011) themselves (trained on two time slices from the Google Books corpus with texts from the 1960's and the 1990's) and the sense-based SCAN model by Frermann and Lapata (2016) (trained on the DATE corpus with texts from the 1960's through the 1990's).[8]

For all our metrics, the *max* across the four time intervals—i.e., identifying the pair of successive intervals where the strongest shift has occurred (cf. end of Section 4.2.4)—is the best performing aggregation strategy. Table 4.2 only shows values obtained with *max* and Euclidean distance for APD, as these are the best performing options.

It is interesting to observe that APD can prove as informative as JSD and ED although it does not depend on the clustering of word occurrences into usage types. Yet, computing usage types offers a powerful tool for analysing lexical change, as we will see in the next section.

[8]Gulordava and Baroni (2011) report Pearson correlation. However, to allow for direct comparison, Frermann and Lapata (2016) computed Spearman correlation for that work (see their footnote 7), which is the value we report.

Frequency difference	0.068
Entropy difference (*max*)	0.278
Jensen-Shannon divergence (*max*)	0.276
Average pairwise distance (*Euclidean*, *max*)	0.285
Gulordava and Baroni (2011)	0.386
Frermann and Lapata (2016)	0.377

Table 4.2: Spearman's correlation coefficients between the gold standard scores in the GEMS dataset and the change scores assigned by our three metrics and a relative frequency baseline. For reference, correlation coefficients reported by previous works using different approaches are also given. All correlations are significant ($p < 0.05$) except for the frequency difference baseline.

4.5 Qualitative analysis

In this section, we provide an in-depth qualitative analysis of the linguistic properties that define usage types and the kinds of lexical semantic change we observe. More quantitative methods (such as taking the top n words with highest JSD, APD and ED and checking, e.g., how many cases of broadening each metric captures) are difficult to operationalise because there exist no well-established formal notions of semantic change types in the linguistic literature (for a discussion of this issue, see Section 3.2.1 or, e.g., Tang et al., 2016). To conduct this analysis, for each GEMS word, we identify the most representative usages in a given usage type cluster by selecting the five closest vectors to the cluster centroid, and take the five corresponding sentences as usage examples.

4.5.1 What do usage types capture?

We first leave the time axis aside and present a synchronic analysis of usage types. The goal is to assess the interpretability and internal coherence of the obtained usage clusters.

We observe that usage types can discriminate between underlying senses of polysemous (and homonymous) words, between literal and figurative usages, and between usages that fulfil different syntactic roles; furthermore, they can single out phrasal collocations as well as named entities.

Polysemy and homonymy. Usage types often encode distinctions between underlying senses of polysemous and homonymous words. For example, the vectors collected for the polysemous word 'curious' are grouped together into two usage types, depending on whether 'curious' is used to describe something that excites attention as odd, novel, or unexpected ('a wonderful and *curious* and unbelievable story') or rather to describe someone who is marked by a desire to

investigate and learn ('*curious* and amazed and innocent'). The same happens, for instance, for the homonymous usages of the word 'coach', which can denote vehicles as well as instructors (see Figure 4.5a for a diachronic view of the usage types).

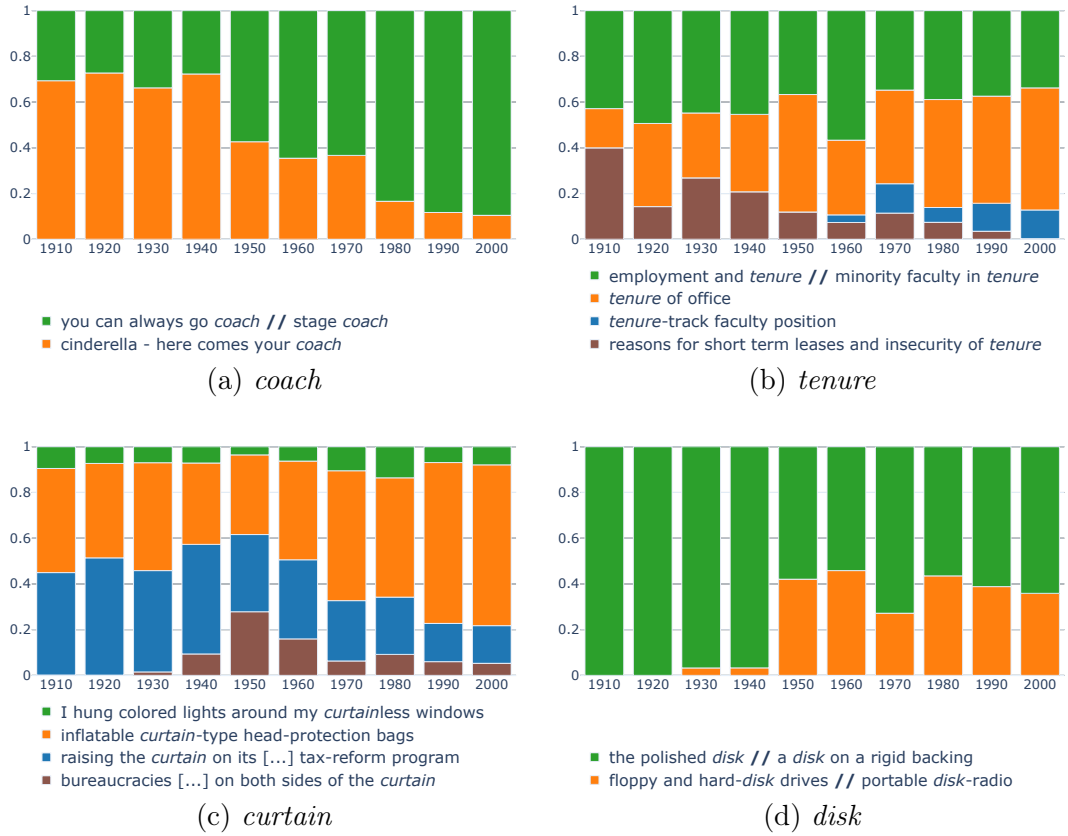
Metaphor and metonymy. In several cases, literal and metaphorical usages are also separated. For example, occurrences of 'curtain' are clustered into four usage types (Figure 4.5c): two of these correspond to a literal interpretation of the word as a hanging piece of cloth ('*curtain*less windows', 'pulled the *curtain* closed') whereas the other two indicate metaphorical interpretations of 'curtain' as any barrier that excludes the free exchange of information or communication ('the *curtain* on the legal war is being raised'). Similarly, we obtain two usage types for 'sphere': one for literal usages that denote a round solid figure ('the *sphere* of the moon'), and the other for metaphorical interpretations of the word as an area of knowledge or activity ('a certain *sphere* of autonomy') as well as metonymical usages that refer to the planet Earth ('land and peoples on the top half of the *sphere*').

Syntactic roles and argument structure. Further distinctions are observed between word usages that fulfil a different syntactic functionality: not only is part-of-speech ambiguity detected (e.g., 'the *cost*-tapered average tariff' vs. '*cost* less to make') but contextualised representations also capture regularities in syntactic argument structures. For example, usages of 'refuse' are clustered into nominal usages ('society's emotional *refuse*', 'the amount of *refuse*'), verbal transitive and intransitive usages ('fall, give up, *refuse*, kick'), as well as verbal usages with infinitive complementation ('*refuse* to go', '*refuse* for the present to sign a treaty').

Collocations and named entities. Specific clusters are also assigned to lexical items that are parts of phrasal collocations (e.g., 'iron *curtain*') or of named entities ('alexander graham *bell*' vs. '*bell*-like whistle').

Other distinctions. Some distinctions are interpretable but unexpected. As an example, the word 'doubt' does not show the default noun-verb separation but rather a distinction between usages in affirmative contexts ('there is still *doubt*', 'the benefit of the *doubt*') and in negative contexts ('there is not a bit of *doubt*', 'beyond a reasonable *doubt*').

Observed errors. For some words, we find that usages which appear to be identical are separated into different usage types. In a handful of cases, this seems due to our experimental setup, which sets the minimum number of clusters to 2 (see Section 4.2.3). This leads to distinct usage types for words such as 'maybe', for which a single type is expected.



(a) *coach* (b) *tenure*

(c) *curtain* (d) *disk*

Figure 4.5: Evolution of usage type distributions in the period 1910–2009, generated with occurrences of *coach*, *tenure*, *curtain* and *disk* in COHA (Davies, 2012). The legends show sample usages per identified usage type.

In other cases, a given interpretation is not identified as an independent type, and its usages appear in more than one cluster. This holds, for example, for the word 'tenure', whose usages in phrases such as '*tenure*-track faculty position' are present in two distinct usage types (see Figure 4.5b).

Finally, we see that in some cases a usage type ends up including two interpretations which arguably should have been distinguished. For example, two of the usage types identified for 'address' are interpretable and coherent: one includes usages in the sense of formal speech and the other one includes verbal usages. The third usage type, however, includes a mix of nominal usages of the word which correspond to different underlying word senses, such as 'disrespectful manners or *address*' and 'network *address*'.

4.5.2 What kinds of change are observed?

We now consider usage types diachronically. Different kinds of change, driven by cultural and technological innovation as well as by historical events, emerge

from a qualitative inspection of usage distributions along the temporal dimension. We describe the most prominent kinds—narrowing and broadening, including metaphorisation—and discuss the extent to which our metrics are able to detect them.

Narrowing. Examination of the dynamics of usage distributions allows us to see that, for a few words, certain usage types disappear or become less common over time (i.e., the interpretation of the word becomes 'narrower', less varied). This is the case, for example, for *coach*, where the frequency decrease of one of the usage types is gradual and caused by technological advances (see Figure 4.5a).

Negative mean ED (see Section 4.2.4) reliably indicates this kind of narrowing. Indeed *coach* is assigned one of the lowest ED score among the GEMS words. In contrast, ED fails to detect the obsolescence of a usage type when new usage types emerge simultaneously (since this may lead to no entropy reduction). This is the case, e.g., for *tenure*. The usage type capturing *tenure* of a landed property becomes obsolete; however, we obtain a positive mean ED caused by the appearance of a new usage type (the third type in Figure 4.5b).

Broadening. For a substantial amount of words, we observe the emergence of new usage types (i.e., a 'broadening' of their use). This may be due to technological advances as well as to specific historical events. As an example, Figure 4.5d shows how, starting from the 1950's and as a result of technological innovation, the word *disk* starts to be used to denote also optical disks while beforehand it referred only to generic flat circular objects.

A special kind of broadening is metaphorisation. As mentioned in Section 4.5.1, the usage types for the word *curtain* include metaphorical interpretations. Figure 4.5c allows us to see when the metaphorical meaning related to the historically charged expression *iron curtain* is acquired. This novel usage type is related to a specific historical period: it emerges between the 1930's and the 1940's, reaches its peak in the 1950's, and remains stably low in frequency starting from the 1970's.

The metrics that best capture broadening are JSD and APD—e.g., *disk* is assigned a high semantic change score by both metrics. Yet, sometimes these metrics generate diverging score rankings. For example, *curtain* yields a rather low APD score due to the low relative frequency of the novel usage (Figure 4.5c). In contrast, even though the novel usage type is not very prominent in some decades, JSD can still discriminate it and measure its development. On the other hand, the word *address*, for which we also observe broadening, is assigned a low score by JSD due to the errors in its usage type assignments pointed out in Section 4.5.1. As APD does not rely on usage types, it is not affected by this issue and does indeed assign a high change score to the word.

Finally, although our metrics help us identify the broadening of a word's

meaning, they cannot capture the type of broadening (i.e., the nature of the emerging interpretations). Detecting metaphorisation, for example, may require inter-cluster comparisons to identify a metaphor's source and target usage types, which we leave to future work.

4.6 Evaluation across languages: The SemEval-2020 shared task

In this second study, we extensively evaluate combinations of architectures, training corpora, and change detection algorithms, using 5 test sets in 4 languages. To this end, we participated in a SemEval-2020 shared task. The SemEval-2020 Shared Task 1 challenged its participants to classify a list of target words into stable or changed (Subtask 1) and/or to rank these words by the degree of their semantic change (Subtask 2) (Schlechtweg et al., 2020). The task is multilingual: it includes four lists of target words, respectively for English, German, Latin, and Swedish (several dozen words each). Each word list is accompanied with two historical corpora of varying size, consisting of texts created in two different time periods. The shared task organisers additionally provided frequency-based and distributional baseline methods.

We participated in Subtask 2 as the **UiO-UvA** team.[9] Our systems are based on two language models, ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019b), and employ three different algorithms to compare contextualised embeddings diachronically (two from the previous study and a novel one). Our *evaluation phase* submission to the shared task ranked 9$^{\text{th}}$ out of 34 participating teams, while our *post-evaluation phase* submission remains as of today the best from those published on the shared task website[10] (but some knowledge of the test sets statistics was needed, see below).

Our main findings are twofold: (i) in 3 out of 5 test sets, ELMo consistently outperforms BERT, while being much faster in training and inference; (ii) cosine similarity of averaged contextualised embeddings and average pairwise distance between contextualised embeddings are the two best performing change detection algorithms, but different test sets show strong preference to either the former or the latter. This preference shows strong correlation with the distribution of gold scores in a test set. While it may indicate that there is a bias in the available test sets, this finding remains yet unexplained.

Our implementations of all the evaluated algorithms are available at `https://github.com/akutuzov/semeval2020`, and the ELMo models we trained can be downloaded from the NLPL vector repository.[11]

[9] We did not specifically focus on the binary Subtask 1; our submission achieved the average accuracy of 0.587 in this track.

[10] `https://competitions.codalab.org/competitions/20948`

[11] `http://vectors.nlpl.eu/repository/`

4.6.1 System overview

As in the previous study, given two time periods t_1, t_2, two corpora C_1, C_2, and a set of target words, we use a neural language model to obtain contextualised embeddings of each occurrence of the target words in C_1 and C_2 and use them to compute a continuous change score. This score indicates the degree of semantic change undergone by a word between t_1 and t_2, and the target words are ranked by its value. We use three change detection algorithms:

1. Inverted cosine similarity over word prototypes (PRT). Given two usage matrices $\mathbf{U}_w^{t_1}, \mathbf{U}_w^{t_2}$, the degree of change of w is calculated as the inverted similarity[12] between the average token embeddings ('prototypes') of all occurrences of w in the two time periods:

$$PRT\left(\mathbf{U}_w^{t_1}, \mathbf{U}_w^{t_2}\right) = \frac{1}{d\left(\frac{\sum_{\mathbf{x}_i \in \mathbf{U}_w^{t_1}} \mathbf{x}_i}{N_w^{t_1}}, \frac{\sum_{\mathbf{x}_j \in \mathbf{U}_w^{t_2}} \mathbf{x}_j}{N_w^{t_2}}\right)} \tag{4.6}$$

where $N_w^{t_1}$ and $N_w^{t_2}$ are the number of occurrences of w in time periods t_1 and t_2, and d is a similarity metric, for which we use cosine similarity. This method is similar to the standard LSCD workflow with *static* embeddings produced by Procrustes-aligned time-specific distributional models (Hamilton et al., 2016), with the only additional step of averaging token embeddings to create a single vector. Since we want the algorithm to produce higher scores for the words which changed more, the inverted value of cosine similarity is used as the prediction.

2. Average pairwise cosine distance between token embeddings (APD). Here, the degree of change of w is measured as the average distance between any two embeddings from different time periods. High APD values indicate stronger semantic change. This is the change metric defined in our previous study; see Equation 4.5 in Section 4.2.4.

3. Jensen-Shannon divergence (JSD). This measure relies on the partitioning of embeddings into clusters of similar word usages and measures the amount of change in the proportions of word usage clusters across time periods (Equation 4.4). A high JSD score indicates a high degree of lexical semantic change. See Section 4.2.4 for more details.

4.6.2 Experimental setup

For each of the 4 languages of the shared task, we train 4 ELMo model variants:

[12] We also tried to use cosine distance $(1 - d)$ instead of inverted cosine similarity, but the results were marginally worse.

i. **Pre-trained**, an ELMo model trained on the respective Wikipedia corpus (English, German, Latin or Swedish);[13]

ii. **Fine-tuned**, the same as Pre-trained but further fine-tuned on the union of the two test corpora;

iii. **Trained on test**, trained only on the union of the two test corpora;

iv. **Incremental**, two models—the first is trained on the first test corpus, and the second is the same model further trained on the second test corpus.

The ELMo models are trained for 3 epochs (except English and Latin **Trained on test** and **Incremental** models, for which we use 5 epochs, due to small test corpora sizes), with an LSTM dimensionality of 2048, a batch size of 192 and 4096 negative samples per batch. All other hyperparameters are left at their default values.[14]

For BERT, we again use the *base* version, with 12 layers and 768 hidden dimensions.[15] For English, German and Swedish, we employ language-specific models; for Latin, we resort to a multilingual one since there is no specific Latin BERT available yet.[16] Given the limited size of the test corpora (in the order of 10^8 word tokens at max), we do not train BERT from scratch and only test the **Pre-trained** and **Fine-tuned** BERT variants. The fine-tuning is done with BERT's standard objective for 2 epochs (English was trained for 5 epochs). We configure BERT's WordPiece tokeniser to never split any occurrences of the target words (some target words are split by default into character sequences) and we add unknown target words to BERT's vocabulary. We perform this step both before fine-tuning and before the extraction of contextualised representations.

At inference time, we use all ELMo and BERT variants to produce contextualised representations of all the occurrences of each target word in the test corpora. For the **Incremental** variant, the representations for the occurrences in each of the two test corpora are produced using the respective model trained on this corpus. The resulting embeddings are of size 12×768 and 3×512 for BERT and ELMo, respectively. We employ three strategies to reduce their dimensionality to that of a single layer: (i) using only the top layer, (ii) averaging all layers, (iii) averaging the last four layers (only for BERT embeddings, as this

[13]The Wikipedia corpora were lemmatised using UDPipe (Straka and Straková, 2017) prior to training.

[14]To train and fine-tune ELMo models, we use the code from `https://github.com/ltgoslo/simple_elmo_training`, which is essentially the reference ELMo implementation updated to the recent TensorFlow versions.

[15]We rely on Hugging Face's implementation of BERT (available at `https://github.com/huggingface/transformers`, version 2.5.0), and follow their model naming conventions: `https://huggingface.co/models`.

[16]*bert-base-uncased*, *bert-base-german-cased*, *af-ai-center/bert-base-swedish-uncased*, *bert-base-multilingual-cased*.

aggregation method was shown to work on par with the all-layers alternative by Devlin et al. (2019b)). Finally, to predict the strength of semantic change of each target word between the two test corpora, we feed the word's contextualised embeddings into the three algorithms of semantic change estimation described in Section 4.6.1. We then compute the Spearman correlation of the estimated change scores with the gold answers. This is the evaluation metric of Subtask 2, and we use it throughout our experiments.

4.6.3 Results

Our submission. In our official shared task submission in the evaluation phase, we used top-layer ELMo embeddings with the cosine similarity change detection algorithm for all languages. English and German ELMo models were trained on the respective Wikipedia corpora. For Swedish and Latin, pre-trained ELMo models were not available, so we trained our own models on the union of the test corpora. This combination of architectures and algorithms was chosen based on our preliminary experiments with the available human-annotated semantic change datasets for English (Gulordava and Baroni, 2011), German (Schlechtweg et al., 2018) and Russian (Fomin et al., 2019). The resulting Spearman correlations were 0.136 for English, 0.695 for German, 0.370 for Latin, and 0.278 for Swedish. With an average score of 0.37, this submission ranked 9$^{\text{th}}$ out of 34 teams in the evaluation phase.

We were aware that the submitted setup was likely sub-optimal as it did not include the **Fine-tuned** model variant. After the official submission deadline (in the post-evaluation phase), we finished training and fine-tuning all of our language models. Their systematic evaluation is the main contribution of this study.

Current results. The average scores of all the tested configurations across 4 languages are given in Table 4.3. This table includes both the results of the configurations we used in the evaluation phase and the results of the configurations we tested after the submission deadline (fine-tuned models). We compare our scores to the organisers' baselines (FD and CNT+CI+CD, as provided by Schlechtweg et al. (2020)) and the classical approach of calculating cosine distance between static CBOW word embeddings (Mikolov et al., 2013). The CBOW models were used in two different flavors: (i) 'incremental', where the C_2 model was initialised with the C_1 weights (Kim et al., 2014), and (ii) 'Procrustes', where the two models were trained independently on C_1 and C_2, and then aligned using the Orthogonal Procrustes transformation (Hamilton et al., 2016). All the training hyperparameters for both ELMo and BERT were fixed to their default values (see Section 4.6.2), we only varied the training corpora and the layers from which embeddings were extracted.

Baselines	Frequency (FD)			-0.083
	Count (CNT+CI+CD)			0.144*
CBOW	Incremental			0.140
	Procrustes			0.392***
Contextualised embeddings		**Top layer**	**Average all layers**	**Average top 4 layers**
		Cosine similarity (PRT)		
BERT	Pre-trained	0.278**	0.233	0.229
	Fine-tuned	0.373**	0.320**	0.338**
ELMo	Pre-trained	0.375**	0.344**	–
	Fine-tuned	0.402**	0.389**	–
	Trained on test	0.370**	0.342**	–
	Incremental	0.114*	0.127	–
		Pairwise distance (APD)		
BERT	Pre-trained	0.237**	0.163*	0.203*
	Fine-tuned	0.363***	0.241**	0.297*
ELMo	Pre-trained	0.296**	0.172*	–
	Fine-tuned	0.405***	**0.406*****	–
	Trained on test	0.338**	0.295***	–
	Incremental	0.126**	-0.001*	–
		Jensen-Shannon divergence (JSD)		
BERT	Pre-trained	0.181*	0.125	0.203*
	Fine-tuned	0.176*	0.223**	0.186**
ELMo	Pre-trained	0.251*	0.196*	–
	Fine-tuned	0.197*	0.156*	–
	Trained on test	0.225*	0.163*	–
	Incremental	-0.037	-0.009	–

Table 4.3: Spearman correlation coefficients for Subtask 2 averaged over four languages. The number of asterisks denotes the number of languages for which the correlation was statistically significant ($p < 0.05$).

Table 4.3 shows that no method achieves statistically significant correlation on all four languages, which attests both to the difficulty of the task and the diversity of the test sets. CBOW Procrustes is a surprisingly strong approach, consistently outperforming the organisers' baselines. Only PRT and APD obtain higher average scores, with fine-tuned ELMo models performing better than the fine-tuned BERT.

Judging only from the average correlation scores, contextualised embeddings do not seem to outshine their static counterparts, especially considering that both ELMo and BERT are more computationally demanding than CBOW. However, closer analysis of per-language results shows that in fact the contextualised approaches outperform the CBOW Procrustes baseline by a large margin for *each* of the shared task test sets. Table 4.4 features the scores obtained by our best-performing methods (PRT and APD with top layer embeddings from fine-tuned

	Algorithm	*English*	*German*	*Latin*	*Swedish*	*GEMS*
CBOW	Incremental	0.210	0.145	0.217	-0.012	**0.424**†
	Procrustes	0.285	0.439†	0.387†	0.458†	0.235†
Fine-tuned contextualised embeddings (top layer)						
ELMo	Cosine similarity (PRT)	0.254	**0.740**†	0.360†	0.252	0.323†
	Average pairwise distance (APD)	**0.605**†	0.560†	-0.113	**0.569**†	0.323†
BERT	Cosine similarity (PRT)	0.225	0.590†	**0.561**†	0.185	0.394†
	Average pairwise distance (APD)	0.546†	0.427†	0.372†	0.254	0.243†

Table 4.4: Spearman correlation per test set for our best methods (post-evaluation phase). † marks statistical significance ($p < 0.05$).

ELMo and BERT) on the individual languages of the shared task. We also report performance on the GEMS ('GEometrical Models of Natural Language Semantics workshop') test set (Gulordava and Baroni, 2011) to enable a comparison with our previous study. The discrepancy between the averaged and the per-language results can be explained by properties of the test sets: APD works best on the English and Swedish sets, while PRT yields the best scores for German and Latin.

Although consistency across languages (3 out of 4) is an important benefit of the CBOW Procrustes approach, with the right choice of APD or PRT, contextualised embeddings can improve Spearman's correlation coefficients by up to 50%. This is not a language-specific property: the English GEMS test set does not behave like the English test set from the shared task. In fact, one can observe 3 groups of test sets with regards to their statistical properties and to the method they favour: group 1 (Latin and German) exhibits rather uniform gold score distributions and prefers PRT; group 2 (English and Swedish) is characterised by more skewed gold score distributions and prefers APD; group 3 (GEMS) is in between, with no clear preference.

Interestingly, the method which produces a more uniform predicted score distribution (APD) works better for the test sets with skewed gold distributions, and the method which produces a more skewed predicted score distribution (PRT) works better for the uniformly distributed test sets (see Figures 4.6 and 4.7). Furthermore, there is perfect negative correlation ($\rho = -1$) between the median gold score of a test set and the performance of the APD algorithm with fine-tuned ELMo models on this test set. The same correlation for the APD performance is not significant but strictly negative. We currently do not have a plausible explanation for this behaviour.

In the bottom part of Figure 4.6, we show how different the 5 test sets are in terms of how the *gold* scores are distributed in them. In some test sets, the gold scores are skewed to the left, while some have a more uniform distribution. The top part of Figure 4.6 shows the distributions of the *predicted* scores produced by the APD and PRT algorithms (with fine-tuned ELMo embeddings).

PRT tends to squeeze the majority of predictions near the lower boundary (no semantic change), with a low median score. In contrast, APD distributes its predictions in a much more uniform way, with a higher median score. Counter-intuitively, skewed gold distributions favour uniform predictions and vice versa. The grouping differences can be quantified with respect to the median gold score (after unit-normalisation). Figure 4.7 shows the dependency of the PRT and APD performance on the median score of the gold test set. The dots here are the performance values of PRT or APD algorithms on different test sets. English and Swedish test sets are in the left part of the plot with the median gold scores of 0.200 and 0.203 correspondingly. German, GEMS and Latin are on the right with 0.266, 0.267 and 0.364 correspondingly. There is a perfect negative Spearman's correlation between the median gold scores of these 5 test sets and the performance of APD semantic change detection algorithm on each of them (with fine-tuned ELMo embeddings).

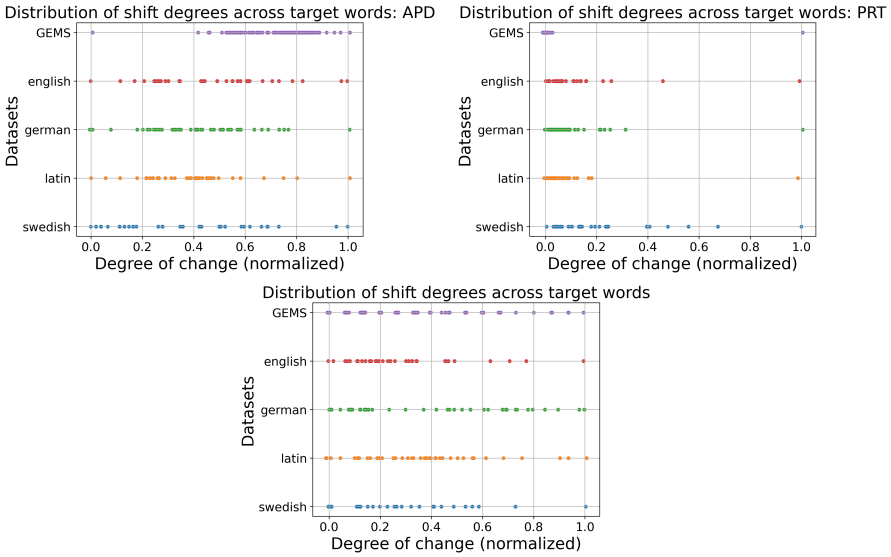


Figure 4.6: Bottom: distribution of semantic change degree in the **gold** data; top left: distribution of scores predicted by the **APD** algorithm; top right: distribution of scores predicted by the **PRT** algorithm.

Furthermore, Table 4.4 supports the previous observation that ELMo models perform better than BERT in the LSCD task. The only test set for which this is not the case is Latin, while on GEMS, ELMo and BERT are on par.[17] One possible explanation is that our ELMo models were pre-trained on lemmatised Wikipedia corpora and thus better fit the test corpora, provided in lemmatised form by the organisers. The BERT models were pre-trained on raw corpora, and

[17]The Latin test corpora are very peculiar: (i) homonyms in them are followed by '#' and the sense identifier, which is not the case for Latin Wikipedia, (ii) the sizes of C_1 and C_2 are very imbalanced, with the latter being 4 times larger than the former.

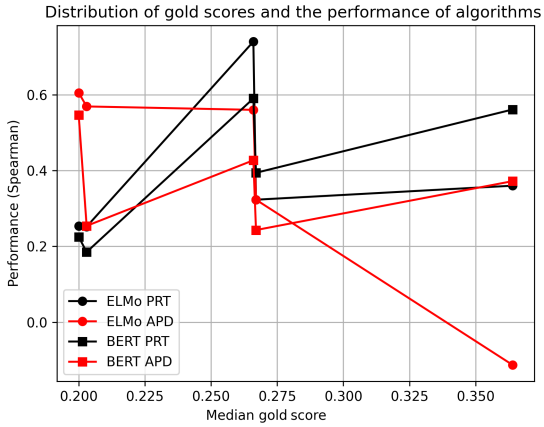


Figure 4.7: Performance of the PRT and APD algorithms depending on the median gold score.

fine-tuning them on lemmatised data proves less successful. This is of course not an advantage of the ELMo architecture *per se*; however, easy and fast training from scratch on the respective Wikipedia corpora for each shared task language was possible only because of much lower computational requirements of ELMo compared to BERT.[18]

In the post-evaluation phase of the shared task, we submitted predictions obtained with the optimal system configurations: fine-tuned ELMo + APD for English and Swedish, fine-tuned ELMo + PRT for German, and fine-tuned BERT + PRT for Latin. This submission reached the average Spearman correlation of 0.618 and, at the time of writing, it is still the best Subtask 2 submission for SemEval-2020 Task 1 (among those publicly available on the shared task website). Of course, the optimal choice of configurations was possible only because we already knew the test data. Still, it is useful for understanding of the real abilities of contextualised embedding-based approaches and the peculiarities of different models and test sets.

4.7 Conclusion

We have introduced a novel approach for the analysis and detection of lexical semantic change in text corpora. To our knowledge, ours is the first work to tackle this problem using neural contextualised word representations and no lexicographic supervision. We have shown that the representations and the detected semantic shifts are aligned to human interpretation, and presented a new

[18]Note that Martinc et al. (2020) report a Spearman correlation of 0.510 on the GEMS dataset using fine-tuned BERT embeddings with Affinity Propagation and JSD. However, we were unable to reproduce these results, even when using the published code.

dataset of human similarity judgements for English word usages which can be used to measure said alignment. Through in-depth qualitative analysis, we have demonstrated that our method allows us to capture a variety of synchronic and diachronic linguistic phenomena.

Participating in the SemEval-2020 shared task, we have experimented with alternative ways of obtaining contextualised representations (using a different language model, fine-tuning, and various layer selection strategies) and our extensive analyses, across four languages, have further confirmed that using contextualised representations to rank words by the degree of their semantic change yields strong correlation with human judgements, outperforming approaches based on static embeddings.

Our approach offers several advantages over previous methods: (i) it does not rely on a fixed number of word senses, (ii) it captures morphosyntactic properties of word usage, and (iii) it allows for a more interpretable quantification of lexical meaning, by enabling the inspection of particular example sentences. Future work could investigate whether usage representations can provide an even finer grained account of lexical meaning and its dynamics, e.g., to automatically discriminate between different types of meaning change.

As of today, the work presented in this chapter has already enabled analyses of variation and change which exploit the expressiveness of contextualised word representations (Kapron-King and Xu, 2021; Lucy and Bamman, 2021; Lucy et al., 2022; Fugikawa et al., 2023, i.a.).

Chapter 5

Contextualised definitions as interpretable word representations

The content of this chapter is based on the following publication:

> Mario Giulianelli, Iris Luden, Raquel Fernández, and Andrey Kutuzov. 2023. Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Mario and Andrey jointly produced the research idea. Mario and Andrey performed the majority of the experiments and wrote a large part of the article. Iris contributed to a smaller part of the experiments and of the writing. Raquel provided advice throughout the project, she reviewed and revised the manuscript. All authors participated in the human evaluation study presented in the paper. The text in this chapter overlaps with that of the original publication.

5.1 Introduction

In the previous chapter, we have seen that lexical semantic change analysis, a task which requires capturing word meaning with its nuanced context-determined modulations, is a good test-bed for lexical representations, and that contextualised word embeddings extracted from pre-trained language models (LMs) are an accurate type of lexical representation. While they can be clustered and manually analysed, however, contextualised embeddings inherently lack in explainability and interpretability due to their subsymbolic nature. This makes the main potential end users of these technologies—historical linguists, lexicographers, and social scientists—still somewhat reluctant to adopt them. Lexicographers, for instance, are not satisfied with detecting that a word has or has not changed its meaning over the last ten years; they want descriptions of old and new senses in human-readable form, possibly accompanied by additional layers of explanation, e.g., specifying the type of semantic change (such as broadening, narrowing, and metaphorisation) the word has undergone.

This study is an attempt to bridge the gap between modern lexical representations and their users. We propose to replace black-box contextualised embeddings produced by large LMs with a new type of interpretable lexical semantic representation: automatically generated *contextualised word definitions* (Gardner et al., 2022). In this paradigm, the usage of the word 'apple' in the sentence 'She tasted a fresh green apple' is represented not with a dense high-dimensional vector but with the context-dependent natural language definition *'edible fruit'*. With an extended case study on lexical semantic change analysis, we show that moving to the more abstract meaning space of definitions allows practitioners to obtain explainable predictions from computational systems, while leading to superior performance on semantic change benchmarks compared to state-of-the-art token-based approaches.

The work presented in this chapter makes the following contributions.

- We show that word definitions automatically generated with a specialised language model, fine-tuned for this purpose, can serve as interpretable representations for polysemous words (Section 5.4). Pairwise usage similarities between contextualised definitions approximate human semantic similarity judgements better than similarities between usage-based word and sentence embeddings.
- We present a method to obtain *word sense representations* by labelling data-driven clusters of word usages with sense definitions, and collect human judgements of definition quality to evaluate these representations (Section 5.5). We find that sense labels produced by retrieving the most prototypical contextualised word definition within a group of usages consistently outperform labels produced by selecting the most prototypical token embedding.

- Using sense labels obtained via definition generation, we create maps that describe diachronic relations between the senses of a target word. We then demonstrate how these *diachronic maps* can be used to explain meaning changes observed in text corpora and to find inconsistencies in data-driven groupings of word usages within existing lexical semantic resources (Section 5.6).

Our code, which can be used to reproduce and build on our experiments, can be found at `https://github.com/ltgoslo/definition_modeling`.

Usage example	**Target word**	**Generated definition**
'about half of the soldiers in our rifle platoons were **draftees** whom we had trained for about six weeks'	**draftee**	*'a person who is being enlisted in the armed forces'*

Table 5.1: An example of a definition generated by our fine-tuned Flan-T5 XL for the word 'draftee' The model is prompted with the usage example, post-fixed with the phrase 'What is the definition of draftee?'

5.2 Data

5.2.1 Datasets of word definitions

To train an NLG system that produces definitions (Section 5.3), we use three datasets containing a human-written definition for each lexicographic sense of a target word, paired with a usage example. The **WordNet** dataset is a collection of word definitions and word usages extracted by Ishiwatari et al. (2019) from the WordNet lexical database (Miller, 1995). The **Oxford** dataset (also known as CHA in prior work) consists of definitions and usage examples collected by Gadetsky et al. (2018) from the Oxford Dictionary. Definitions are written by experts and usage examples are in British English. The **CoDWoE** dataset (Mickus et al., 2022) is based on definitions and examples extracted from Wiktionary.[1] It is a multilingual corpus, of which we use the English portion. Table 5.2 reports the main statistics of these datasets. Further statistics, e.g., on the size of the different splits, are provided by Huang et al. (2021) as well as in Appendix A.1.[2]

[1] `https://www.wiktionary.org`

[2] A definition dataset could be also be extracted from the SemCor corpus (Miller et al., 1993). However, we do not anticipate it will contribute much to training or evaluation since SemCor does not contain any new definitions with respect to WordNet: only more examples for the same word-definition pairs. This can be investigated in future work.

Dataset	Entries	Lemmas	Ratio	Usage length	Definition length
WordNet	15,657	8,938	1.75	4.80 ± 3.43	6.64 ± 3.77
Oxford	122,318	36,767	3.33	16.73 ± 9.53	11.01 ± 6.96
CoDWoE	63,596	36,068	2.44	24.04 ± 21.05	11.78 ± 8.03

Table 5.2: Main statistics of the datasets of definitions. Ratio is the *sense-lemma* ratio: the number of entries over the number of lemmas.

5.2.2 Diachronic word usage graphs

We showcase interpretable word usage (Section 5.4) and sense representations (Section 5.5 and 5.6) using a dataset where target lemmas are represented with diachronic word usage graphs (DWUGs, Schlechtweg et al., 2021). A DWUG is a weighted, undirected graph, where nodes represent target usages (word occurrences within a sentence or discourse context) and edge weights represent the semantic proximity of a pair of usages. DWUGs are the result of a multi-round incremental human annotation process, with annotators asked to judge the semantic relatedness of pairs of word usages on a 4-point scale (similar to our annotation process in the previous chapter). Based on these pairwise judgements, word usages are then grouped into usage clusters (a data-driven approximation of *word senses*) using a variation of correlation clustering (Bansal et al., 2004; Schlechtweg et al., 2020).

DWUGs are currently available in seven languages.[3] Here, we use the English graphs, which consist of usage sentences sampled from the Clean Corpus of Historical American English (Davies, 2012; Alatrash et al., 2020) and belonging to two time periods: 1810-1860 and 1960-2010. There are 46 usage graphs for English, corresponding to 40 nouns and 6 verbs annotated by a total of 9 annotators. Each target lemma has received on average 189 judgements, 2 for each usage pair. Figure 5.1 shows an example of a DWUG, with colours denoting usage clusters (i.e., data-driven senses). The 'blue' and 'orange' clusters belong almost entirely to different time periods: a new sense of the word has emerged. We show how our approach helps explain such cases of semantic change in Section 5.6.

5.3 Definition generation

Our formulation of the *definition generation* task is as follows: given a target word w and an example usage s (i.e., a sentence containing an occurrence of w), generate a natural language definition d that is grammatical, fluent, and faithful to the meaning of the target word w as used in the example usage s. This corresponds to the sequence-to-sequence task formulation discussed in Section 3.3.

[3] `https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/wugs/`

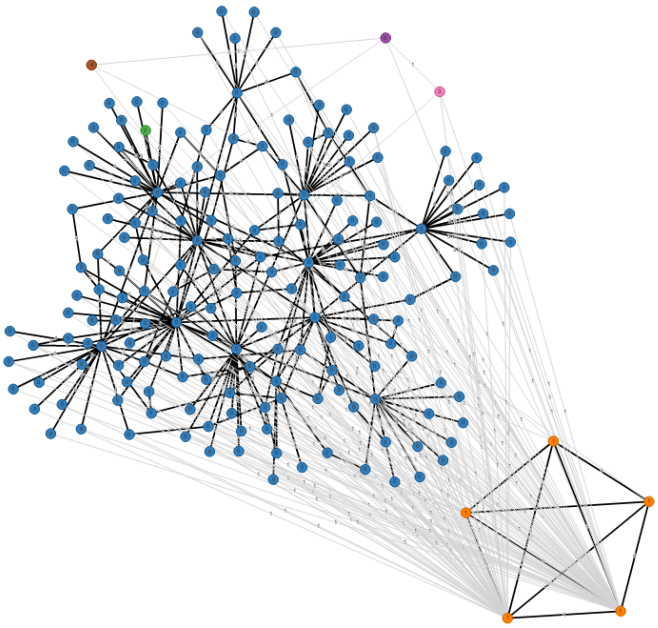


Figure 5.1: Diachronic word usage graph for the English word 'lass' (Schlechtweg et al., 2021).

A *definition generator* is a language process $\mathbf{G}: \mathscr{V} \times \mathscr{L} \rightarrow \mathscr{L}$ that maps words and example usages to natural language definitions. $\mathbf{G}: (w, s) \mapsto d$. In this study, $\mathscr{L}$ is the English language and $\mathscr{V}$ is the English vocabulary, which we further restrict to lists of selected target words as available in definition datasets (see Section 5.2). As a generator, we use Flan-T5 (Chung et al., 2022), a version of the T5 encoder-decoder Transformer (Raffel et al., 2020) fine-tuned on 1.8K tasks phrased as instructions and collected from almost 500 NLP datasets. Flan-T5 is not trained specifically on definition generation but thanks to its massive multi-task instruction fine-tuning, the model exhibits strong generalisation to unseen tasks. Therefore, we expect it to produce high-quality definitions. We extensively test three variants of Flan-T5 of different size (and compare them to vanilla T5 models in Appendix A.3); based on our results, we recommend using the largest fine-tuned Flan-T5 model whenever possible.

To obtain definitions from Flan-T5, we use natural language prompts consisting of an example usage preceded or followed by a question or instruction. For example: 's `What is the definition of` w?'. The concatenated usage example and prompt are provided as input to Flan-T5, which conditionally generates definitions (Table 5.1 shows an example instance). This is a simpler workflow in comparison to prior work (Bevilacqua et al., 2020; Almeman and Espinosa Anke, 2022) where inputs are encoded as 'target word - context' pairs. We choose greedy search with target word filtering as a simple, parameter-free decoding strategy. Stochastic decoding algorithms can be investigated in future work.

Prompt selection. In preliminary experiments, we used the pre-trained Flan-T5 Base model (250M parameters) to select a definition generation prompt among 8 alternative verbalisations. Appending the question *'What is the definition of w?'* to the usage example consistently yielded the best scores (further details in Appendix A.2). We use this prompt for all further experiments.

5.3.1 Evaluating generated definitions

Before using its definitions to construct an interpretable semantic space—the main goal of this study—we perform a series of experiments to validate Flan-T5 as a definition generator. We use the target lemmas and usage examples from the corpora of definitions presented in Section 5.2, conditionally generate definitions with Flan-T5, and then compare them to the gold definitions in the corpora using reference-based NLG evaluation metrics. We report SacreBLEU and ROUGE-L, which measure surface form overlap, as well as BERT-F1, which is sensitive to the reference and candidate's semantics. As mentioned in the background chapter (Section 3.3), reference-based metrics are not flawless, yet designing and validating a reference-free metric for the definition generation task is beyond the scope of this study. We will later resort to correlations with human judgements and expert human evaluation to assess the quality of generated definitions.

We evaluate Flan-T5 XL as a definition generator (3B parameters) by conducting four generalisation tests: 1) zero shot (task shift), 2) in distribution, 3) hard domain shift, and 4) soft domain shift. The tests are defined following the GenBench generalisation taxonomy (Hupkes et al., 2023). An evaluation card which clarifies the nature of the tests is shown in Table 5.3.[4]Results obtained using another language model, T5 (Raffel et al., 2020), are presented in Appendix A.3.

We use the generalisation tests to choose a model to be deployed in further experiments. For reference, we report the BLEU score of the definition generator by Huang et al. (2021); ROUGE-L and BERT-F1 are not reported in their paper.

Zero-shot. We directly evaluate Flan-T5 XL on the WordNet and Oxford test sets, without any fine-tuning nor in-context learning.[5] Table 5.4 shows low BLEU and ROUGE-L scores but rather high BERT-F1. Overall, the model does not exhibit consistent task understanding (e.g., it generates *'skepticism'* as a definition for 'healthy' as in the phrase 'healthy skepticism'). A qualitative inspection, however, reveals that the generated definitions can still be often informative (e.g., *'a workweek that is longer than the regular workweek'* is informative with respect to the meaning of 'overtime' although the ground truth definition is *'beyond the*

[4]See `https://genbench.org/eval_cards`. In-distribution tests are not included as they do not present any shift between the training and test data distributions Hupkes et al. (2023).

[5]We only condition generation on the usage examples and the task prompt. We do *not* provide full instances (i.e., usage examples, task prompts, and definitions) in the context, as one would do in a few-shot setup.

Motivation			
Practical	*Cognitive*	*Intrinsic*	*Fairness*
□ △ ◯			
Generalisation type			
Compo- sitional	*Structural*	*Cross Task*	*Cross Language*	*Cross Domain*	*Robust- ness*
		□		△ ◯	
Shift type			
Covariate	*Label*	*Full*	*Assumed*
△ ◯		□	
Shift source			
Naturally occurring	*Partitioned natural*	*Generated shift*	*Fully generated*
□ △ ◯			
Shift locus			
Train–test	*Finetune train–test*	*Pretrain–train*	*Pretrain–test*
	△ ◯		□

Table 5.3: Evaluation card for the generalisation tests performed on definition generators. The setups are: zero-shot (□), hard domain shift (△), and soft domain shift (◯). In-distribution tests are not included as they do not include any shift between the training and test data distributions.

regular time'). The two surface overlap metrics cannot capture this, but the relatively high BERT-F1 confirms that the semantic content of generations is largely appropriate. There are indeed also many good zero-shot definitions—for example, *'intense'* for 'fervent' as in 'the fervent heat', or *'a conversation'* for 'discussion' in 'we had a good discussion'.

In distribution. We fine-tune Flan-T5 XL on one corpus of definitions at a time, and test it on a held-out set from that same corpus (except CoDWoE which does not provide train-test split). The quality of the definitions increases substantially with fine-tuning, in terms of both their lexical and semantic overlap with gold definitions (Table 5.4). We find significantly higher scores on Oxford, which may be due to the larger size of its training split and to the quality of the WordNet examples, which sometimes are not sufficiently informative (Almeman and Espinosa Anke, 2022).

Hard domain shift. We fine-tune Flan-T5 XL on WordNet and test it on Oxford, and vice versa. These tests allow us to assess the model's sensitivity to the peculiarities of the training dataset. A model that has properly learned to generate definitions should be robust to this kind of domain shift. The quality of the definitions of Oxford lemmas generated with the model fine-tuned on WordNet (see the Oxford column in Table 5.4) is lower than for the model fine-tuned on

		WordNet			***Oxford***		
Model	Test	BLEU	ROUGE-L	BERT-F1	BLEU	ROUGE-L	BERT-F1
Huang et al. (2021)	*Unknown*	32.72	-	-	**26.52**	-	-
Flan-T5 XL	Zero-shot (task shift)	2.70	12.72	86.72	2.88	16.20	86.52
Flan-T5 XL	In distribution	11.49	28.96	88.90	16.61	36.27	89.40
Flan-T5 XL	Hard domain shift	29.55	48.17	91.39	8.37	25.06	87.56
Flan-T5 XL	Soft domain shift	**32.81**	**52.21**	**92.16**	18.69	**38.72**	**89.75**

Table 5.4: Results of the definition generation experiments.

Oxford itself (same column, see row 'In distribution'). Instead, for out-of-domain WordNet definitions, all metrics surprisingly indicate higher quality than for in-distribution tests (WordNet column). Taken together, our results so far suggest that the quality of a fine-tuned model depends more on the amount of the training data, and on the quality of the usage examples in the dataset, than on whether the test data is drawn from the same dataset.

Soft domain shift. We finally fine-tune Flan-T5 XL on a collection of all three definition datasets: WordNet, Oxford, and CoDWoE. Our previous results hint towards the model's preference for more training examples, so we expect this setup to achieve the highest scores regardless of the soft shift between training and test data. Indeed, on WordNet, our fine-tuned model marginally surpasses the state-of-the-art upper bound in terms of BLEU score (Table 5.4), and it achieves the highest scores on the other metrics. Oxford definitions generated with this model are instead still below Huang et al.'s upper bound; this may be due to Oxford being generally more difficult to model than WordNet, perhaps because of longer definitions and usages (see Figures A.1-A.2 in Appendix A.1). We consider the observed model performance sufficient for the purposes of our experiments, in particular in view of the higher efficiency of fine-tuned Flan-T5 with respect to the three-module system of Huang et al. (2021). We therefore use this model throughout the rest of our study.

The Flan-T5 models fine-tuned for definition generation are publicly available through the Hugging Face model hub.[6]

5.4 Definitions are interpretable word representations

We propose considering the abstract meaning space of definitions as a representational space for lexical meaning. Definitions fulfil important general desiderata

[6]Model names: ltg/flan-t5-definition-en-base, ltg/flan-t5-definition-en-large, ltg/flan-t5-definition-en-xl.

Method	Cosine	SacreBLEU	METEOR
Token embeddings	0.141	-	-
Sentence embeddings	0.114	-	-
FLAN-T5 XL Zero-shot	0.188	0.041	0.083
FLAN-T5 XXL Zero-shot	0.206	0.045	0.092
FLAN-T5 base (fine-tuned)	0.221	0.078	0.077
FLAN-T5 XL (fine-tuned)	**0.264**	**0.108**	**0.117**

Table 5.5: Correlations with pairwise similarity judgements by humans.

for word representations: they are human-interpretable and they can be used for quantitative comparisons between word usages (i.e., by judging the distance between pairs of definition strings). We put the *definition space* to test by applying it to the task of semantic change analysis, which requires capturing word meaning at a fine-grained level, distinguishing word senses based on usage contexts. We use our fine-tuned Flan-T5 models (XL and other sizes) to generate definitions for all usages of the 46 target words annotated in the English DWUGs (ca. 200 usages per word; see Section 5.2.2).[7] These definitions (an example is provided in Table 5.1) specify a diachronic semantic space.

5.4.1 Correlation with human judgements

We construct word usage graphs for each lemma in the English DWUGs: we take usages as nodes and assign weights to edges by measuring pairwise similarity between usage-dependent definitions. We compute the similarity between pairs of definitions using two overlap-based metrics, SacreBLEU and METEOR, as well as the cosine similarity between sentence-embedded definitions. We then compare our graphs against the gold DWUGs, where edges between usage pairs are weighted with human judgements of semantic similarity, by computing the Spearman's correlation between human similarity judgements and similarity scores obtained for pairs of generated definitions. We compare our results to DWUGs constructed based on two additional types of usage-based representations: *sentence* embeddings obtained directly for usage examples, and contextualised *token* embeddings. Sentence embeddings (for both definitions and usage examples) are SBERT representations (Reimers and Gurevych, 2019) extracted with mean-pooling from the last layer of a DistilRoBERTa LM fine-tuned for semantic similarity comparisons.[8] For tokens, we extract the last-layer representations

[7]The training datasets used in Section 5.3 contain nouns, verbs, adjectives and adverbs. The English DWUGs contain only nouns and verbs.

[8]DistilRoBERTa (`sentence-transformers/all- distilRoBERTa-v1`) is the second best model as reported in the official S-BERT documentation at the time of publication (`https:`

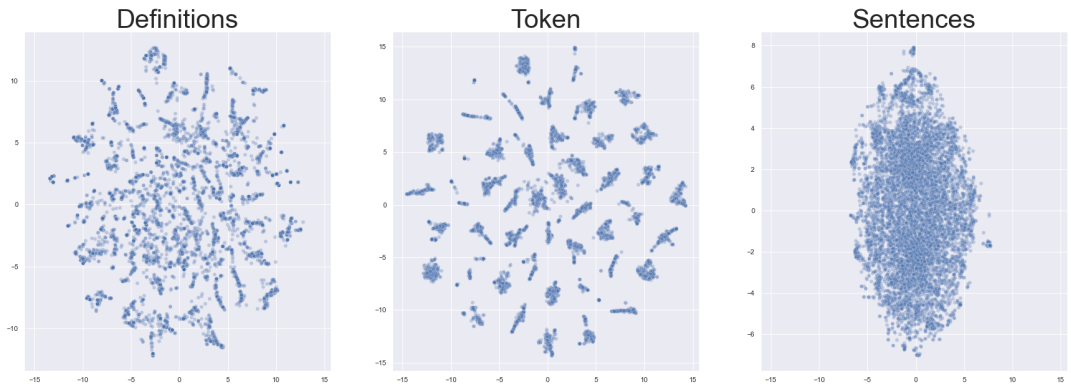


Figure 5.2: T-SNE projection of each embedding space, DistilRoBERTa model.

of a RoBERTa-large model (Liu et al., 2019) which correspond to subtokens of the target word following the procedure presented in the previous chapter, and then use mean-pooling to obtain a single vector. While we report string-overlap similarities for definitions, these are not defined for numerical vectors, and thus similarities for example sentences and tokens are obtained with cosine only.

Pairwise similarities between definitions approximate human similarity judgements far better than similarities between example sentence and word embeddings (Table 5.5) indicating that definitions are a more accurate approximation of contextualised lexical meaning. The results also show that similarity between definitions is best captured by their embeddings, rather than by overlap-based metrics such as SacreBLEU and METEOR.

5.4.2 Definition embedding space

We now examine the *definition embedding space* (the high-dimensional semantic space defined by sentence-embedded definitions), to identify properties that make it more expressive than usage-based spaces. Figure 5.2 shows the T-SNE projections of the DistilRoBERTa embeddings of all lemmas in the English DWUGs, for the three types of representation presented earlier: generated definitions, tokens, and example sentences.[9] The definition spaces exhibit characteristics that are more similar to a *token* embedding space than an example *sentence* embedding space, with usages of the same lemma represented by relatively close-knit clusters of definition embeddings. This suggests that definition embeddings, as expected, represent the meaning of a word in context (similar to token embeddings), rather than the meaning of the whole usage example sentence in which the target word occurs.

`//www.sbert.net/docs/pretrained_models.html`). For a negligible accuracy reduction, it captures longer context sizes and is ca. 50% smaller and faster than the model that ranks first.

[9] T-SNE projections for RoBERTa-large are in Appendix A.6.

Model	Representation	Variance	Std	k	Silhouette ↑	Sep. ↑	Coh. ↓	Ratio ↑
RoBERTa-large	Sentence	0.014	0.117	2.0	0.111	0.285	0.012	23.2
	Token	0.034	0.183	3.8	0.173	**0.868**	0.027	**32.4**
	Definitions	0.006	0.080	20.6	**0.335**	0.057	**0.003**	22.3
DistilRoBERTa	Sentence	0.597	0.772	2.1	0.046	4.907	0.578	8.5
	Token	0.477	0.687	2.5	0.121	**8.599**	0.427	20.1
	Definitions	0.509	0.756	19.7	**0.355**	5.559	**0.228**	**24.4**

Table 5.6: Variance, standard deviation, optimal k, silhouette score, separation score, cohesion score, and the separation-cohesion ratio for each embedding space; average over all target words.

For each target word, we also measure (i) the variability in each embedding space and (ii) the inter-cluster and intra-cluster dispersion (Caliński and Harabasz, 1974) obtained when clustering each space using k-means. This allows us to quantitatively appreciate properties exhibited by data-driven usage clusters that are obtained from different representation types. To cluster the embedding spaces, we experiment with values of $k \in [2, 25]$, and select the k which maximises the silhouette score (Rousseeuw, 1987). Our results are summarised in Table 5.6. While, on average, token spaces exhibit higher inter-cluster dispersion (indicating better cluster separation), the clusters in the definition spaces have on average the lowest intra-cluster dispersion, indicating that they are more cohesive than the clusters in the token and example sentence spaces. These findings persist for the gold clusters determined by the English DWUGs (Table A.4, Appendix A.6).

In sum, this analysis shows that definition embedding spaces are generally suitable to distinguish different types of word usage. In the next section, we will show how they can indeed be used to characterise word senses.

5.5 Labelling word senses with definitions

For generated definitions to be useful in practice, they need to be able to distinguish word senses. For example (ignoring diachronic differences and singleton clusters), there are three main senses of the word 'word' in its DWUG, which we manually label as: (1) *'words of language'*, (2) *'a rumour'*, and (3) *'an oath'*. Manual inspection of the generated definitions indicates that they are indeed sense-aware:

1. *'A communication, a message'*, *'The text of a book, play, movie'*, etc.
2. *'Information passed on, usually by one person to another'*, *'communication by spoken or written communication'*, etc.
3. *'An oath'*, *'a pronouncement'*, etc.

But let's again put ourselves in the shoes of a historical linguist. Sense clusters are now impractically represented with multitudes of contextualised definitions. Cluster (1) for 'word', e.g., features 190 usages, and one must read through all of them and extrapolate—all to formulate a definition that covers the whole sense cluster (a *sense label*). We now show how DWUGs can be automatically augmented with generated sense labels, vastly improving their usability.

5.5.1 Selecting sense labels

From n definitions, generated for n word usages belonging to the same DWUG cluster, we use the most prototypical one as the *sense label*—with the aim of reflecting the meaning of the majority of usages in the cluster. We represent all definitions with their sentence embeddings (cf. Section 5.4.1) and select as prototypical the definition whose embedding is most similar to the average of all embeddings in the cluster. Clusters with less than 3 usages are ignored as, for these, prototypicality is ill-defined. As a sanity check, these are the sense labels obtained by this method for the DWUG clusters of 'word'; they correspond well to the sense descriptions provided earlier.

1. *'A single spoken or written utterance'*
2. *'Information; news; reports'*
3. *'A promise, vow or statement'*

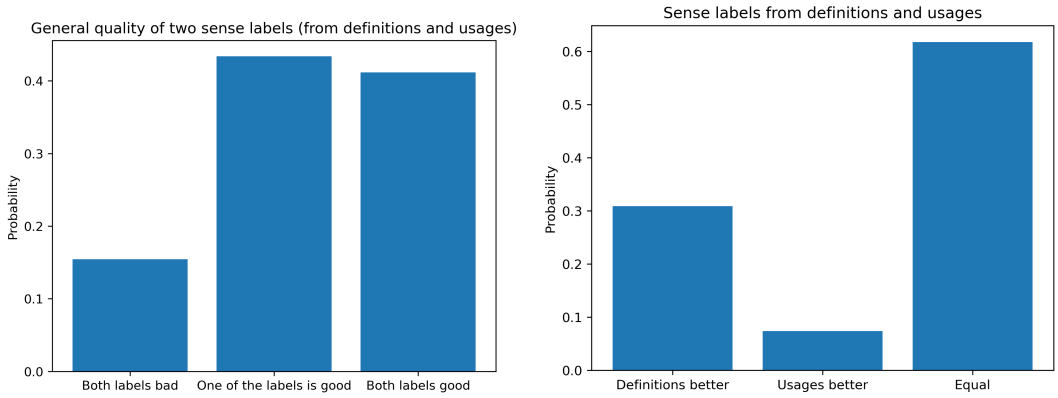
We compare these sense labels to labels obtained by generating a definition for the most prototypical *usage* (as judged by its token embedding), rather than taking the most prototypical *definition*, and we evaluate both types of senses labels using human judgements. Examples of labels can be found in Appendix A.4.

5.5.2 Human evaluation

Five human annotators (fluent English speakers) were asked to evaluate the quality of sense labels for each cluster in the English DWUGs, 136 in total. Each cluster was accompanied by the target word, two labels (from definitions and from usages) and five example usages randomly sampled from the DWUG. The annotators could select one of six judgements to indicate overall quality of the labels and their relative ranking. After a reconciliation round, the categorical judgements were aggregated via majority voting. Krippendorff's α inter-rater agreement is 0.35 on the original data and 0.45 when the categories are reduced to four. Full guidelines are reported in Appendix A.5.

There exist no established procedures for the collection of human quality judgements of automatically generated word sense labels. The closest efforts we are aware of are those in Noraset et al. (2017), who ask annotators to rank definitions generated by two systems, providing as reference the gold dictionary

definitions. In our case, (i) generations are for word senses rather than lemmas, (ii) we are interested not only in rankings but also in judgements of 'sufficient quality', (iii) dictionary definitions are not available for the DWUG senses; instead (iv) we provide annotators with usage examples, which are crucial for informed judgements of sense definitions.



(a) (b)

Figure 5.3: Human evaluation results: general quality of generated sense labels (a) and comparison between sense labels produced from definitions vs. usages.

Results. Figure 5.3 shows the results of the human evaluation. We find that our prototypicality-based sense labelling strategy is overall reliable. Only for 15% of the clusters, annotators indicate that neither of the labels is satisfactory (Figure 5.3a). When comparing definition-based and usage-based labels (Figure 5.3b), the former were found to be better in 31% of the cases, while the latter in only 7%. In the rest of the cases, the two methods are judged as equal. We also analysed how often the labels produced by each method were found to be acceptable. Definition-based labels were of sufficient quality in 80% of the instances, while for usage-based labels this is only true for 68% of the cases.

In sum, prototypical definitions reflect sense meanings better than definitions of prototypical usage examples. We believe this is because definitions are more abstract and robust to contextual noise: if the underlying sense is similar, the same definition can be assigned to very different usages. This approach takes the best of both worlds. The produced representations are data-driven, but at the same time they are human-readable and naturally explanatory. In the next section, we demonstrate how automatically generated definition-based sense labels can be used to explain semantic change observed in diachronic text corpora.

5.6 Explaining semantic change with sense labels

Word senses in DWUGs are collections of example usages and they are only labelled with numerical identifiers. This does not allow users to easily grasp the meaning trajectories of the words they are interested in studying. Using sense labels extracted from generated definitions, we can produce a fully human-readable *sense dynamics map*—i.e., an automatically annotated version of a DWUG which displays synchronic and diachronic relations between senses (e.g, senses transitioning one into another, splitting from another sense, or two senses merging into one).

Given a target word, its original DWUG, and its semi-automatic sense clusters, we start by assigning a definition label to each cluster, as described in Section 5.5. Then, we divide each cluster into two sub-clusters, corresponding to time periods 1 and 2 (for example, sub-cluster c_1^2 contains all usages from cluster 1 occurring in time period 2).[10] We compute pairwise cosine similarities between the sentence embeddings of the labels (their 'definition embeddings'), thereby producing a fully connected graph where nodes are sub-clusters and edges are weighted with sense label similarities. Most edges have very low weight, but some sub-cluster pairs are unusually similar, hinting at a possible relation between the corresponding senses. We detect these outlier pairs by inspecting the distribution of pairwise similarities for values with z-score higher than 1 (similarities more than 1 standard deviation away from the mean similarity). Sub-cluster pairs connected with such edges form a *sense dynamics map*.

As an example, the noun 'record' has only one sense in time period 1 but it acquires two new senses in time period 2 (Figure 5.4; as before, we ignore clusters with less than 3 usages). The sense clusters defined by DWUGs are anonymous collection of usages, but with the assigned sense labels (also shown in Figure 5.4) they can be turned into an explanation of the observed semantic shift:

- A novel sense 2 of 'record' in time period 2 (*'A phonograph or gramophone cylinder containing an audio recording.'*) is probably an offshoot of a stable sense 0 present in both time periods (*'A document or other means of providing information about past events.'*).

It becomes now clear that sense 2 stems from the older general sense 0 of 'record'—arguably representing a case of narrowing (Bloomfield, 1933)—while the second new sense (1: *'the highest score or other achievement in the game'*) is not related to the others (or at least much further related) and can thus be considered as independent.

[10]Note that the labels are still time-agnostic: that is, sub-clusters c_1^1 and c_1^2 have the same label. This is done for simplicity and because of data scarcity, but in the future we plan to experiment with time-dependent labels as well. We use two time periods as only two periods are available in Schlechtweg et al.'s English DWUGs (Schlechtweg et al., 2021), but the same procedure can be executed on multi-period datasets.

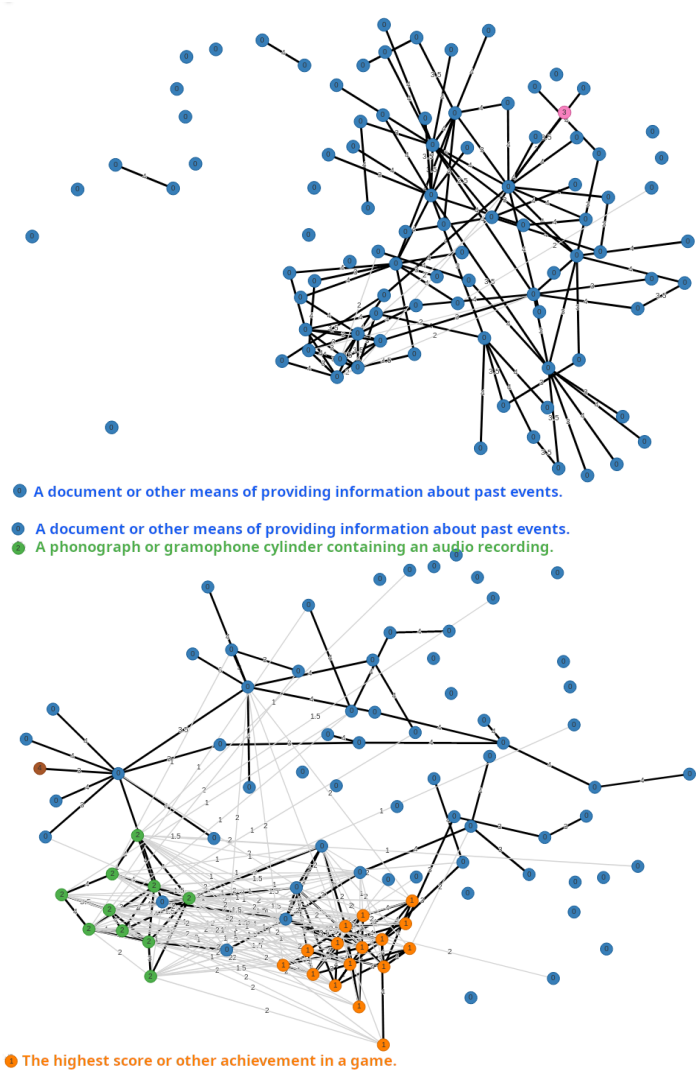


Figure 5.4: Diachronic word usage graphs for 'record' (Schlechtweg et al., 2021) with sense definitions generated using our proposed procedure (Section 5.5). Left: time period 1 (1810-1860); right: time period 2 (1960-2010). Colours correspond to data-driven senses, as annotated in the original DWUGs.

Sense dynamics maps can also help in tracing potentially incorrect or inconsistent clustering in DWUGs. For instance, if different sense clusters are assigned identical definition labels, then it is likely that both clusters correspond to the same sense and that the clustering is thus erroneous. Using our automatically produced sense dynamics maps, DWUGs can be improved and enriched semi-automatically.

An interesting case in which different sense clusters are assigned identical definition labels is the word 'chef'. Usage examples from sense clusters c_2 and c_3 for the word 'chef' are as follows:

- c_2: '*He boasted of having been a* ***chef*** *de brigade in the republican armies of France*', '*Morrel has received a regiment, and Joliette is* ***Chef*** *d'Escadron of Spahis*', '*as major-general and* ***chef*** *d'escadron, during the pleasure of our glorious monarch Louis le Grand*'

- c_3: '*That brave general added to his rank of* ***chef*** *de brigade that of adjutant general*', '*I frequently saw Mehevi and several other* ***chefs*** *and warriors of note take part*'

A user can safely accept the suggestion of our system to consider these two clusters as one sense.[11] Another insightful case is 'ball'. Although none of its sense labels are identical, its sense cluster c_0 is very close to cluster c_2 (similarity of 0.70), while c_2 is close to c_3 (similarity of 0.53); all three senses persist throughout both time periods, with sense 3 declining in frequency. The generated definitions for the 'ball' clusters are: 0: *'A sphere or other object used as the object of a hit'* (the largest cluster), 2: *'A round solid projectile, such as is used in shooting'*, and 3: *'A bullet'*. This case demonstrates that similarity relations are not transitive: the similarity between c_0 and c_3 is only 0.50, below our outlier threshold value. This is in part caused by inconsistent DWUG clustering: while the majority of usages in c_2^1 are about firearm projectiles, c_2^2 contains mentions of golf balls and ball point pens. This shifts sense 2 from *'bullet'* to *'round solid projectile'*, making it closer to sense 0 (general spheres) than it should be. Ideally, all the *'bullet'* usages from c_2 should have ended up in c_3, with the rest joining the general sense 0.

Besides suggesting fixes to the DWUG clustering, the observed non-transitivity also describes a potential (not necessarily diachronic) meaning trajectory of 'ball': from any spherical object, to spherical objects used as projectiles, and then to any projectiles (like bullets), independent of their form. Our generated sense labels and their similarities help users analyse this phenomenon in an easier and considerably faster way than by manually inspecting all examples for these senses.

[11] Note that *'A commander'* practically disappeared as a word sense in the 20th century, replaced by *'a professional cook, usually in a restaurant'*.

5.7 Conclusion

We propose to consider automatically generated contextualised word definitions as a type of lexical representation, similar to traditional word embeddings. While generated definitions have been already shown to be effective for word sense disambiguation (Bevilacqua et al., 2020), our study puts this into a broader perspective and demonstrates that modern language models like Flan-T5 (Chung et al., 2022) are sufficiently mature to produce robust and accurate definitions in a simple prompting setup. The generated definitions outperform traditional token embeddings in word-in-context similarity judgements while being naturally interpretable.

We apply definition-based lexical representations to semantic change analysis and show that our approach can be used to trace word sense dynamics over time. Operating in the space of human-readable definitions makes such analyses much more interesting and actionable for linguists and lexicographers—who look for explanations, not numbers. At the same time, we believe the 'definitions as representations' paradigm can also be used for other NLP tasks in the area of lexical semantics, such as word sense induction, idiom detection, and metaphor interpretation.

Our experiments with diachronic sense modelling are still preliminary and mostly qualitative. The cases shown in Section 5.6 are hand-picked examples, demonstrating the potential of using generated definitions for explainable semantic change detection and improving LSCD datasets. It is important to evaluate systematically how well our predictions correspond to the judgements of (expert) humans. Once further evidence is gathered, other promising applications include tracing cases of semantic narrowing or widening over time (Bloomfield, 1933) by analysing the variability of contextualised definitions in different time periods and by making cluster labels time-dependent. Both directions will require extensive human annotation, and we leave them for future work.

Part Two

Utterance Comprehension

The main focus of Part 1 was on the usage and interpretation of individual words. Chapters 4 and 5 presented two ways of using neural language models to obtain abstract word representations. The resulting representations can be used for linguistic analysis and, at the same time, inspecting them reveals the ability of language models to accurately capture word meaning. While the ability to model word meaning is fundamental to any account of language use, it is not sufficient for a model to offer a complete picture of the much more complex forms of linguistic behaviour humans are capable to produce. Most of the times, human linguistic signals consist of sequences of words, embedded in a wider linguistic context than the current sentence. These sequences, which we will refer to as *utterances*, make up conversations, books, theatre plays, and movie scripts. They are arguably the main unit of language use because, unlike words (which, incidentally, are not a linguistically universal concept), they can describe situations and states of the world (i.e., entities and the relations between them) and thus they can be used to express non-trivial communicative intents.

Utterances will be the focus of Part 2 and Part 3 of this thesis. We will study them in relation to the linguistic context that precedes their occurrence, their *discourse context*, as this is essential to understanding their communicative function and, with that, most aspects of their production and comprehension in humans. This will not simply be an extension of the analyses presented in Part 1 to the more complex meaning space of utterances. While extending contextualised meaning representations to this more complex space is interesting and valuable, it is perhaps more exciting, and arguably more informative about human language use, to study the mechanisms by which humans produce and comprehend such complex signals—rather than, so to say, the 'result' of their processing.

Part 2, in particular, will explore ways of using neural language models to mimic the processes involved in the comprehension of utterances and to study how these determine speakers' audience-aware strategies of language production. Combining information theory and psycholinguistics, we will model utterance comprehension as a process in which humans form expectations about the next complex linguistic unit, and where the amount of processing necessary to make sense of the unit is related to its unpredictability. This, in turn, affects speakers' selection and formulation of utterances, since sequences of words that require excessive processing effort in order to be comprehended are unlikely to be communicatively felicitous.

Due to production and perception errors, differences between individuals, and other sources of uncertainty, language use can be understood as information exchange through a noisy channel. Speakers are sensitive to the properties of the channel in two ways (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989).

On the one hand, they try to reduce the processing effort of the addressee. For example, in the absence of established discourse context, speakers can produce utterances that are easier to process in order to minimise the chance of transmission error. On the other hand, speakers try to reduce their own production effort. For example, given a fixed amount of information that they intend to transmit, speakers can take the risk of producing more concise utterances that are less costly from the production point of view, and expect the addressee to exploit the utterance context for interpretation. Effective and efficient information exchange under these two competing pressures can be modelled using the tools of Information Theory (Shannon, 1948). Indeed, information-theoretic models have offered successful accounts of speech perception (Jelinek et al., 1975; Clayards et al., 2008), reading (Keller, 2004; Demberg and Keller, 2008; Levy et al., 2009), sentence interpretation (Levy, 2008b; Gibson et al., 2013), and turn taking (Dethlefs et al., 2016), providing psycholinguistic evidence that the information content (or *surprisal*) of linguistic signals is related to comprehension effort.

The most efficient way of dealing with the efficiency and effectiveness pressures, according to Information Theory, is to transmit information at a constant surprisal rate (Genzel and Charniak, 2002), making linguistic choices that reduce fluctuations in the density of the information transmitted. Evidence for the principle of uniform information density (UID; Jaeger and Levy, 2007; Jaeger, 2010) has been found at many levels of language production: speakers tend to reduce the duration of more predictable sounds (Aylett and Turk, 2004, 2006; Bell et al., 2003; Demberg et al., 2012); they tend to drop sentential material within more predictable scenarios (Jaeger and Levy, 2007; Jaeger, 2010; Frank and Jaeger, 2008); in spoken dialogue they are more likely to overlap at turn transitions when surprisal is low (Dethlefs et al., 2016); and the rate at which they transmit information in texts is uniform (Genzel and Charniak, 2002, 2003; Qian and Jaeger, 2011).

In Chapter 6, I will present the central notion of surprisal and its relation to processing effort, as well as a method to obtain surprisal estimates from neural language models. In Chapter 7, I will use surprisal estimates to test the entropy rate constancy (Genzel and Charniak, 2002) and uniform information density (Jaeger and Levy, 2007) hypotheses in text and dialogue. Chapter 8 will zoom in on the information structure of utterances in dialogue, trying to reconcile findings in the preceding chapter with theories of rational and efficient use of the communication channel.

Chapter 6

Estimating surprisal with neural language models

The content of this chapter is based on the following publications:

> Mario Giulianelli and Raquel Fernández. 2021. Analysing Human Strategies of Information Transmission as a Function of Discourse Context. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 647–660, Online. Association for Computational Linguistics.
>
> Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is Information Density Uniform in Task-Oriented Dialogues?. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

For both studies, Mario produced the research idea with input and guidance from Raquel. Mario performed the experiments with Arabella and Raquel's supervision. Mario wrote the article, Arabella and Raquel reviewed and contributed to the writing. The text in this chapter overlaps with that of the original publications.

6.1 Background

To ensure communicative success, humans monitor the effect of their utterances on their audience's comprehension. Taking into consideration the audience's *processing effort*, in particular, is one of the main forms of audience-awareness speakers are known to possess (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989; Frank and Goodman, 2012a; Levy, 2018). In many cases, language production choices can be explained in terms of efficient strategies to manage fluctuations of expected effort throughout communication episodes, such as texts and dialogues.

But what is processing effort? Processing effort is a phycholinguistic construct developed as a measure of the expenditure of information processing resources required for the perception and cognition of linguistic signals. It is typically measured in terms of neural responses or other forms of comprehension behaviour such as reading times and eye fixation duration. According to expectation-based psycholinguistic theories of language processing (Hale, 2001; Levy, 2008a), processing effort is strongly related to the predictability of upcoming linguistic signals given their context of occurrence. Computational estimates of predictability offer a convenient solution to the quantification of processing effort because they allow measuring predictability without access to the human brain or behaviour.

A classic and empirically successful operationalisation of predictability relies on the information-theoretic notion of *surprisal*, or *information content* (Shannon, 1948). Surprisal is an alternative way of expressing the probability of a particular event occurring from a random variable, and it can be interpreted as quantifying the degree of unexpectedness of (or surprise for) a particular outcome. It is often also referred to as *information density*. Quantifications of predictability in terms of surprisal have not only been shown to be a strong predictor of neural and behavioural measures of processing effort in perception (Jelinek et al., 1975; Clayards et al., 2008), reading (Keller, 2004; Demberg and Keller, 2008; Levy et al., 2009; Monsalve et al., 2012; Goodkind and Bicknell, 2018a; van Schijndel and Linzen, 2018), and sentence comprehension (Levy, 2008b; Gibson et al., 2013; Shain et al., 2020)—they have also been used to explain a wide variety of phenomena in language production, from phonology (Aylett and Turk, 2004, 2006; Bell et al., 2003; Demberg et al., 2012) and syntax (Jaeger and Levy, 2007; Frank and Jaeger, 2008; Jaeger, 2010) to discourse (Genzel and Charniak, 2002, 2003; Qian and Jaeger, 2011) and dialogue (Dethlefs et al., 2016; Xu and Reitter, 2018).

In lack of computational models that could capture predictability of complex linguistic signals within long contextual sequences, surprisal used to be estimated independently of its discourse context. This chapter presents a method for the estimation of utterance surprisal as a function of discourse context, which relies on autoregressive Transformer language models.

6.2 Method

In this section, I define the main information theoretic measures that will be used throughout Part 2 and describe the computational models that produce empirical estimates thereof. We will take utterances as the basic unit of information transmission in line with prior work on text and dialogue (Genzel and Charniak, 2002, 2003; Doyle and Frank, 2015a,b; Qian and Jaeger, 2011; Xu and Reitter, 2018) and obtain estimates of their surprisal both when considered out of and in their discourse context. To further investigate information structure *within utterances*, in Chapter 8, I will generalise our definition of surprisal such that it can be applied to any sequence of tokens.

6.2.1 Measuring surprisal

The surprisal of a linguistic signal, in our case an utterance S, is the negative logarithm of the probability of S. It quantifies the degree of unexpectedness of—or surprise for—S. This quantity is also called the Shannon information content:

$$H(S) = -\log_2 P(S) \tag{6.1}$$

In this formulation, the surprisal of an utterance measures how unexpected the utterance is if processed out of context. However, because utterances always appear within some discourse, their true surprisal is necessarily modulated by the informativeness of their context. The availability of contextual cues (e.g., the topic of the text, references to the main entities in the discourse, the writing style) alters the expectations of the audience over upcoming linguistic signals and, in most cases, makes utterances less surprising and less effortful to process.

The contextualised surprisal of an utterance, too, can be estimated as the Shannon information content, but this time using the negative logarithm of the *conditional probability* of the utterance given its context C:

$$H(S|C) = -\log_2 P(S|C) \tag{6.2}$$

According to the classic information-theoretic model of communication (Shannon, 1948) and under a rationality assumption for speakers' linguistic choices, this quantity is hypothesised to remain constant (Genzel and Charniak, 2002) or uniform (Jaeger and Levy, 2007; Jaeger, 2010) throughout discourse. In the introduction to Part 2, I have referred to the respective hypotheses as Entropy Rate Constancy (ERC) and Uniform Information Density (UID).

Although both hypotheses generate predictions about contextualised surprisal, previous studies have tried to confirm or disprove them by relying only on estimates of decontextualised surprisal, due to the lack of suitable computational models. To motivate this simplification, they have relied on the assumption that an increase in the amount of available discourse context always corresponds to an

increase in context informativeness (Genzel and Charniak, 2002, 2003): i.e., for example, that three sentences are more informative to predict the fourth sentence in a text than two sentences are to predict the third sentence. The operationalisation of this assumption requires rewriting the contextualised surprisal of an utterance as the difference between the decontextualised surprisal and the mutual information between the next utterance and the context:

$$H(S|C) \equiv H(S) - I(S;C) \tag{6.3}$$

As the relevant context is built up, $I(S;C)$ is assumed to increase. So for the ERC and UID hypotheses to hold—i.e., for $H(S|C)$ to remain constant or uniform in Equation 6.3—the decontextualised surprisal $H(S)$ must increase. In prior work, an increase in $H(S)$ was therefore considered sufficient evidence in favour of the principles (Genzel and Charniak, 2002, 2003).

In our studies in Part 2, we do not assume an increase in $I(S;C)$ and estimate both the decontextualised and the contextualised surprisal of an utterance. This allows us to directly test the ERC and UID hypotheses and to measure the true informativeness of context.

6.2.2 Definitions

The surprisal of a word choice w_i is the negative logarithm of the corresponding word probability, conditioned on the utterance context $S_{:w_i}$ (i.e., the words that precede w_i in utterance S) and on the discourse context C:

$$H(w_i|S_{:w_i}, l) = -\log_2 P(w_i|S_{:w_i}, C) \tag{6.4}$$

While we take the utterance as the basic unit of information transmission, there are multiple reasons to measure surprisal at the word level. It makes it possible to estimate the surprisal of subsequences of tokens within utterances (as we will see in Chapter 8, it is compatible with current neural language model architectures (which output probabilities of tokens, rather than utterances), and it is an approach taken in most previous work (e.g., Genzel and Charniak, 2002; Xu and Reitter, 2018; Meister et al., 2021), thus enhancing the comparability of our results to existing findings.

The *decontextualised surprisal* of an utterance is computed by averaging over the negative logarithms of all word probabilities, conditioned only on the preceding words in the utterance context:

$$H(S) = -\frac{1}{|S|}\sum_{w_i \in S} \log_2 P(w_i|w_1, ..., w_{i-1}) \tag{6.5}$$

The *contextualised surprisal* of an utterance is computed as the average per-word negative probability, conditioned on the preceding words in the utterance as well

Pos.	Utterance	$H(S)$	$H(S\|C)$
1	Stanislav Ovcharenko, who represents the Soviet airline Aeroflot here, has some visions that are wild even by the current standards of perestroika.	5.44	5.44
2	In his office overlooking the runway of Shannon Airport, Mr. Ovcharenko enthusiastically throws out what he calls "just ideas":	6.53	5.61
3	First, he suggests, GPA Group Ltd., the international aircraft leasing company based in Ireland, could lease some of its Boeing jetliners to the Soviet airline.	6.10	5.82

Table 6.1: The first three paragraphs of a Penn Treebank article (document id: 36), annotated with their positions within the article (Pos.) and surprisal estimates.

as on the entire relevant discourse context:

$$H(S|C) = -\frac{1}{|S|}\sum_{w_i \in S} \log_2 P(w_i|w_1, ..., w_{i-1}, C) \tag{6.6}$$

Context informativeness is computed as the difference between the previous two quantities:

$$I(S;C) \equiv H(S) - H(S|C) \tag{6.7}$$

Section 6.4 explains how empirical estimates of the quantities above are obtained from neural language models, and Section 6.5 contains an analysis and validation of our surprisal estimates. First, though, I will present the corpora analysed throughout Part 2 of this thesis, on which language models are fine-tuned and of which surprisal estimates are computed.

6.3 Data

We analyse the surprisal of written and spoken English in text and in dialogue. Excerpts from our corpora of written texts and from our written dialogue corpora are shown in Tables 6.1 and 6.2; further excerpts from the corpora can be found in Appendix B.1.

Penn Treebank. The Penn Treebank corpus[1] (Mitchell et al., 1999) contains 2,499 English newspaper articles from the Wall Street Journal. We follow the

[1] `https://catalog.ldc.upenn.edu/LDC99T42`

Pos.	Id.	Utterance	$H(S)$	$H(S\|C)$
1	B	Hi. Two women with bagels?	5.61	5.61
2	A	nope	4.18	4.22
3	A	guy with a beard and big pizza	4.95	4.77
4	B	No. A woman and child in dimly lit room	5.24	5.02
5	A	yep she has a green jacket on	5.46	5.21
6	A	a wood table with empty beer bottles on it	4.42	4.55
7	B	Yes.	4.86	4.91
8	A	ok ready	7.13	7.64
9	B	Done	11.85	11.30
10	A	k go	10.32	10.57

Table 6.2: The first three utterances of a PhotoBook dialogue (dialogue id: 2037), annotated with utterance positions (Pos.), speaker identifier (Id.), and surprisal estimates.

data splits used by Genzel and Charniak (2002, 2003) and divide the corpus into a training set (sections 0–20) and a test set (sections 21–24). We will use this corpus to replicate the findings of Genzel and Charniak (2002) using surprisal estimates from a neural (rather than n-gram) language model.

PhotoBook. The PhotoBook corpus[2] (Haber et al., 2019) contains 2,500 English task-oriented dialogues between two participants who interact via written chat. The task is set up as game with 5 rounds. In every round, each dialogue participant is shown a set of six images which partially overlap with the set shown to their partner. The images change in each round, but a subset reappears, which elicits re-descriptions of images that have already been referred to in the dialogue. The goal of the game is to discover which images are common to both participants. We split these dialogues into a 70% training set (games 0-1751) and a 30% test set (games 1752-2501).

Map Task. The MapTask corpus[3] (Anderson et al., 1991) contains 128 transcribed spoken dialogues consisting of an instruction giver directing an instruction follower to navigate to a point on a map. The participants cannot see each other's map and their respective maps may contain slightly different landmarks. We randomly split the dialogues into a 70% training set and a 30% test set.

[2] `https://dmg-photobook.github.io`
[3] `http://groups.inf.ed.ac.uk/maptask`

Spoken BNC. The Spoken British National Corpus[4] (Love et al., 2017) is a dataset of transcribed open-domain spoken dialogues containing 1,251 contemporary British English conversations, collected in a range of real-life contexts. To be consistent with PhotoBook, Map Task, and previous work (Vega and Ward, 2009; Xu and Reitter, 2018), we select the dialogues that feature only two speakers. We then randomly split these 622 dialogues into a 70% training and a 30% test set.

6.4 Experimental setup

We compute surprisal using GPT-2 (Radford et al., 2019), a pre-trained autoregressive Transformer language model, which allows us to obtain more accurate probability estimates than the n-gram models used in previous work (Genzel and Charniak, 2002, 2003; Doyle and Frank, 2015a,b; Qian and Jaeger, 2011; Xu and Reitter, 2018) as well as to include discourse context in the computation. We rely on HuggingFace's implementation of GPT-2 with default tokenizers and parameters (Wolf et al., 2020) and to adapt the language model to the idiosyncrasies of different types of language use, we finetune it separately on the 70% training split of a given target corpus under analysis. As shown in Table 6.3, finetuning yields a substantial reduction in the model's perplexity. More information on model parameters and the finetuning procedure can be found in Appendix B.2. We use the finetuned language models to estimate decontextualised and contextualised surprisal (Equation 6.5 and 6.6) of the 30% held-out portion of a given corpus.

6.4.1 Fixed context window

We use the language model's context window up to its maximum size (1024 tokens for GPT-2). This means that once the position of an utterance in a document is relatively high (starting to count from the first utterance in the document), the entire window is filled and earlier portions of the context are systematically tossed out. Therefore, the language model cannot exploit long-distance relations involving information present in earlier portions of the discourse that fall outside this window. To ensure that the $H(S|C)$ estimates are unbiased for high utterance positions, we determine, for each corpus c, the first utterance position pos^{c}_{1024} where the sum of context length average and standard deviation across documents is 1024. Our experiments are then executed on all utterances with position smaller or equal to pos^{c}_{1024}.[5].

[4] `http://www.natcorp.ox.ac.uk`

[5] We have tried to substitute GPT-2 with the Transformer-XL language model (Dai et al., 2019) because of its unlimited context window size. In spite of its larger window, however, Transformer-XL yields higher perplexity than GPT-2 on all corpora, hence we decided to use GPT-2. Further reasons to discard Transformer-XL are discussed in Appendix B.2.1

	Pre-trained	Finetuned
Penn Treebank	28.03	21.89
PhotoBook	43.42	14.93
Map Task	880.63	48.36
Spoken BNC	66.47	8.69

Table 6.3: Word-level perplexity of the GPT-2 models on 30% held-out portions of the corpora.

6.4.2 Control runs

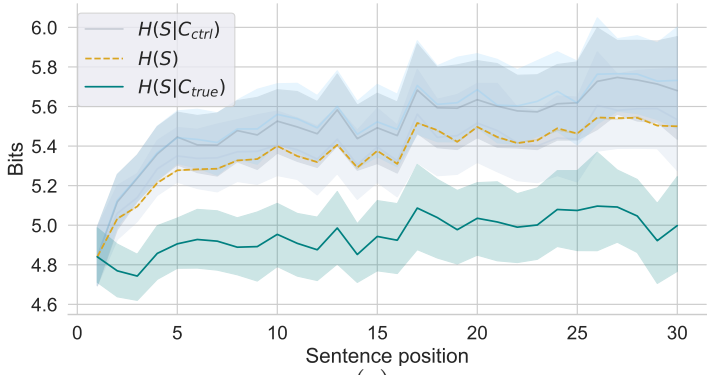
Deep learning models are known to exploit peculiarities of the data distribution that humans would not find relevant. In this case, in particular, we are concerned that our language model may be able to make use of irrelevant contextual features to produce more accurate (but less generalisable) $P(S|C)$ predictions. This would lead to an artificial decrease in $H(S|C)$. To control for this eventuality, we obtain $H(S|C)$ estimates for a given utterance using 3 control contexts, following the same procedure described previously for the true context. We randomly sample one control context from the target corpus and two from a corpus with the same modality (i.e., never mixing monologue and dialogue). This ensures that the control contexts are truly independent with respect to the target utterance (e.g., with respect to topic, referents, and style). The length of the control contexts is always equal to the number of tokens in the true context.

6.5 Analysis of language model estimates

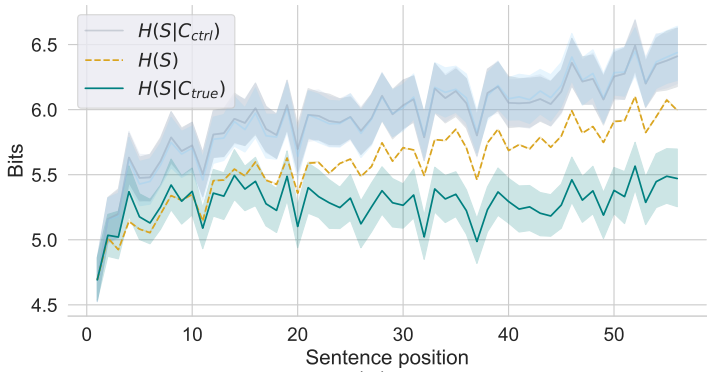
In this section, we report the estimates and patterns of utterance surprisal directly computed with the finetuned GPT-2 language models for three corpora, one for each modality: Penn Treebank, PhotoBook, and Spoken BNC. Recall that we are directly estimating both $H(S)$ and $H(S|C)$ from data, in contrast to previous work, where $H(S|C)$ is never computed empirically.

Before using the $H(S|C)$ estimates to test the ERC and UID hypotheses, we validate them by comparison with those obtained using random control contexts (see Section 6.4.2). If the language model relies on irrelevant contextual features, we would expect the estimates obtained with the true context to be virtually indistinguishable from those obtained with random contexts. This would mean that our $H(S|C)$ estimates are not reliable. In contrast, if the model does effectively exploit the actual context to estimate utterance surprisal, we should see a clear difference between the true $H(S|C)$ estimates and those obtained in the control runs.

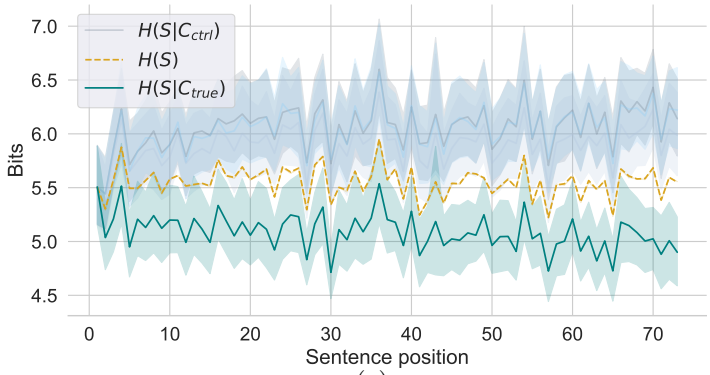
As can be seen in Figure 6.1, true and control trends start diverging from



(a)
Penn Treebank



(b)
PhotoBook



(c)
Spoken BNC

Figure 6.1: Contextualised surprisal estimates with true and random control contexts. Bootstrapped 95% confidence bands. We also show the mean $H(S)$ values for reference (confidence bands will be visible in Figure 7.1).

utterance position 2, as desirable. Control contexts produce a positive shift in the magnitude of $H(S|C)$ in all corpora: processing an utterance S in a random context is always harder than processing it in its true context. Moreover, because the control contexts are incoherent with respect to S, they cause $H(S|C)$ to be higher than $H(S)$: processing an utterance S in an incoherent context is harder than processing it with no context.[6] We also notice that while the magnitude of $H(S|C)$ depends on the veracity of the contexts, its fluctuations are largely determined by $H(S)$. This is particularly true for the control trends of $H(S|C)$, whose slope, too, is determined by $H(S)$. This behaviour is worth further exploration in future work, as it can reveal inherent misalignment in human vs. model language processing.

In sum, the $H(S|C)$ trends computed with the true data differ from the trends obtained with control runs according to reasonable expectations: the true $H(S|C)$ estimates are lower, and the control estimates higher, than the $H(S)$ estimates. This attests to the validity of our empirical estimates of utterance surprisal. Before using these estimates to test the ERC and UID hypotheses (Chapter 7), we also replicate prior results obtained by Genzel and Charniak (2002).

6.6 Replication study: Surprisal constancy in newspaper articles

To validate our estimates of utterance surprisal, we replicate Genzel and Charniak's (2002; 2003) and Keller's (2004) studies on the Wall Street Journal articles (sections 0–24) of the Penn Treebank. We use GPT-2 finetuned on sections 0-20 (see Table 6.3 for its perplexity on sections 21-24). In the original studies, the authors measure the correlation between the position of sentences within newspaper articles—as well as within paragraphs—and sentence surprisal, as measured using n-gram language models. As mentioned above, in Section 6.2.1, these studies assume that $I(X_i; C_i|L_i)$ increases as discourse context is built up, and they test whether the locally conditioned surprisal $H(X_i|L_i)$, too, increases throughout articles and paragraphs.

In our replication study, we take both entire articles and paragraphs as structural units and count sentence positions from the beginning of the relevant unit. Linear mixed effect models show a significant positive effect of sentence position on surprisal both within articles ($\beta = 1.65 \times 10^{-2}, p < 0.001$) and within paragraphs ($\beta = 1.53 \times 10^{-2}, p < 0.01$). To reproduce the original experimental setting, we further train an n-gram language model with interpolated Kneser-Ney smoothing (with $n \in (2, 3, 4, 5)$ and with discount values $d \in (0.1, 0.2, \ldots, 0.9)$)

[6] In Figure 6.1a and, partially, in 6.1c, we can see that one of the control runs is closer to $H(S)$; for this run the contexts are sampled from the target corpus (see Section 6.4.2) and appear to be less harmful for the language model estimates.

and select the configuration with the lowest perplexity on the test set: a 3-gram model with a discount value of 0.8. In line with previous work, we find a positive Kendall's rank correlation[7] between utterance position and information, as measured with the n-gram model as well as with the Transformers (additional results can be found in Appendix B.3). The original results are therefore replicated.

Together with our analysis of language model estimates using control runs, this positive replication study builds confidence in the validity of our surprisal estimates. In the next two chapters, we will use the estimates to study human language production strategies.

[7]Our data consist of multiple measurements for each utterance position (one for each document), thus causing a large number of ties (i.e., multiple entries with the same utterance position but different entropy estimates). We choose Kendall's test because it deals with ties better than other correlation tests such as Spearman's or Pearson's.

Chapter 7

Utterance surprisal as a function of discourse context

The content of this chapter is based on the following publications:

> Mario Giulianelli and Raquel Fernández. 2021. Analysing Human Strategies of Information Transmission as a Function of Discourse Context. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 647–660, Online. Association for Computational Linguistics.
>
> Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is Information Density Uniform in Task-Oriented Dialogues?. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

For both studies, Mario produced the research idea with input and guidance from Raquel. Mario performed the experiments with Arabella and Raquel's supervision. Mario wrote the article, Arabella and Raquel reviewed and contributed to the writing. The text in this chapter overlaps with that of the original publications.

7.1 Introduction

The previous chapter presented surprisal as measure of predictability, its relation to processing effort in psycholinguistic theories and, in turn, to language comprehension and production behaviour, as well as methods to obtain empirical estimates of surprisal from neural language models. We have also discussed how surprisal is expected to vary as a function of discourse context according to a classic information-theoretic model of communication (6.2.1): i.e., it should remain constant, or at least uniform, as discourse develops. In the current chapter, we test this hypothesis extensively in texts and dialogues.

The surprisal of an utterance $H(S)$—i.e., a measure of the effort it takes to process it out of context—and the informativeness of its discourse context $I(S;C)$ are hypothesised to be related. According to the principle of Entropy Rate Constancy (ERC; Genzel and Charniak, 2002), as discourse develops, these two quantities increase at a similar rate; thus, the difference between them—i.e., the effort that it takes to process an utterance *in context*—remains constant over the course of a discourse: $H(S|C) \equiv H(S) - I(S;C)$. A slight relaxation of this prediction is that the surprisal of an utterance in context remains uniform, rather than constant. This second prediction follows from the principle of Uniform Information Density (UID; Jaeger and Levy, 2007; Jaeger, 2010), according to which speakers make rational linguistic choices that avoid peaks in the rate of the information transmitted. Evidence in favour of these principles has been found in texts (Genzel and Charniak, 2002, 2003; Qian and Jaeger, 2011) and, under certain conditions, in conversations (Vega and Ward, 2009; Doyle and Frank, 2015a,b; Xu and Reitter, 2018). However, these studies base their conclusions only on estimates of the decontextualised surprisal $H(S)$: under the assumption that a larger context is always more informative, an increase in $H(S)$ suffices as an indication that the ERC and UID principles hold.

In this study, we dispose of the assumption that context informativeness increases constantly within a discourse, and we test whether the ERC and UID principles hold using, for the first time, direct estimates of the contextualised surprisal $H(S|C)$ of an utterance and thus of the informativity of its linguistic context $I(S;C)$. We use a pre-trained Transformer-based language model, which allows us to obtain more accurate probability estimates than the n-gram models used in previous studies as well to explicitly condition the estimates on discourse context. We loosely follow Genzel and Charniak's procedure (2002; 2003), of which we have already shown an example in the previous chapter, when we replicated their seminal experiments on newspaper articles. In addition, we apply the analysis to open-domain spoken dialogues and to written task-oriented dialogues to test the ERC and UID principles in interactive settings.

Our proposed operationalisation, described in detail in Section 6.2 of the previous chapter, allows us to test whether the increase in decontextualised surprisal observed in earlier work corresponds to an increase in context informativeness,

or whether speakers simply change their information transmission rate over time. Furthermore, this approach allows us to differentiate, for the first time, between the ERC and the UID predictions at the level of discourse.

7.2 Surprisal throughout texts and dialogues: Constancy vs. uniformity

The ERC and UID principles hypothesise that both $H(S)$ and $I(S;C)$ will increase with the position of S within a discourse, and that as a result, $H(S|C)$ will remain stable. This is expressed in Equation 6.3, repeated here for convenience:

$$H(S|C) \equiv H(S) - I(S;C) \tag{7.1}$$

In the following experiments, we investigate whether this is indeed the case in a corpus of written texts (Penn Treebank; Mitchell et al., 1999), a corpus of written task-oriented dialogue (PhotoBook; Haber et al., 2019), and a corpus of open-domain dialogue (Spoken BNC; Love et al., 2017). The three corpora are introduced in the previous chapter, in Section 6.3.

We estimate decontextualised surprisal $H(S)$ (Equation 6.5) and contextualised surprisal $H(S|C)$ (Equation 6.6) with a neural language model, GPT-2, as described in Section 6.2, and using Equation 6.7, we compute context informativeness $I(S;C)$. In the first experiment, we test the constancy hypothesis. In the second, we compare constancy to uniformity as descriptors of utterance surprisal trends over discourse.

7.2.1 Experiment 1: Is surprisal constant?

In Experiment 1, we test whether the positive effect of utterance position on decontextualised surprisal observed in earlier work (e.g., Genzel and Charniak, 2002, 2003; Xu and Reitter, 2018) corresponds to a comparable increase in context informativeness. Following Qian and Jaeger (2011) and Xu and Reitter (2018), we fit linear mixed effect models using the logarithm of the decontextualised surprisal $H(S)$ as our response variable and the logarithm of utterance position as predictor, with a random intercept grouped by distinct documents or dialogues. Because utterance length is known to have an effect on surprisal estimates (Keller, 2004), we include the logarithm of length as an additional predictor. Our models also have a document-specific random slope for utterance position and utterance length to capture cross-document variation (Barr et al., 2013). We repeat the same procedure to also fit models using the logarithm of the contextualised surprisal $H(S|C)$, and the context informativeness $I(S;C)$ as response variables.

The results of the linear mixed effect models are summarised in Table 7.1; a full report of the results is shown in Table B.7 (Appendix B.4). What follows is a discussion of each of the information-theoretic measures in turn.

	$H(S)$	$H(S\|C)$	$I(S;C)$
Penn Treebank	$\beta = 2.94 \times 10^{-2}, p < 0.001$	$\beta = 0.23 \times 10^{-2}, p > 0.05$	$\beta = 12.08 \times 10^{-2}, p < 0.001$
PhotoBook	$\beta = 4.07 \times 10^{-2}, p < 0.001$	$\beta = -1.63 \times 10^{-2}, p < 0.001$	$\beta = 27.94 \times 10^{-2}, p < 0.001$
Spoken BNC	$\beta = -0.05 \times 10^{-2}, p > 0.05$	$\beta = -2.89 \times 10^{-2}, p < 0.001$	$\beta = 6.31 \times 10^{-2}, p < 0.001$

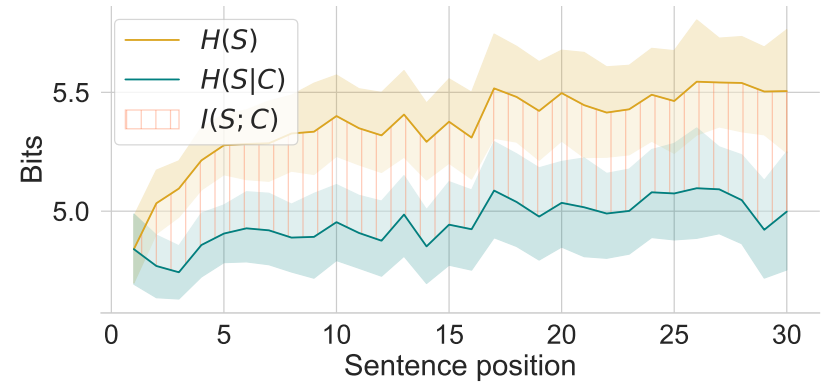
Table 7.1: Coefficients of linear mixed effect models using the logarithm of $H(S)$, $H(S|C)$, and $I(S;C)$ as response variables. The logarithms of utterance position and length are the predictors and they are both assigned a per-document random slope. The models also include a per-document random intercept.

Decontextualised surprisal $H(S)$. Decontextualised surprisal significantly increases with utterance position in Penn Treebank and in PhotoBook. Its rate of increase is relatively low, as indicated by the coefficients of our linear mixed effect model. In Spoken BNC, there is no effect of utterance position on $H(S)$.

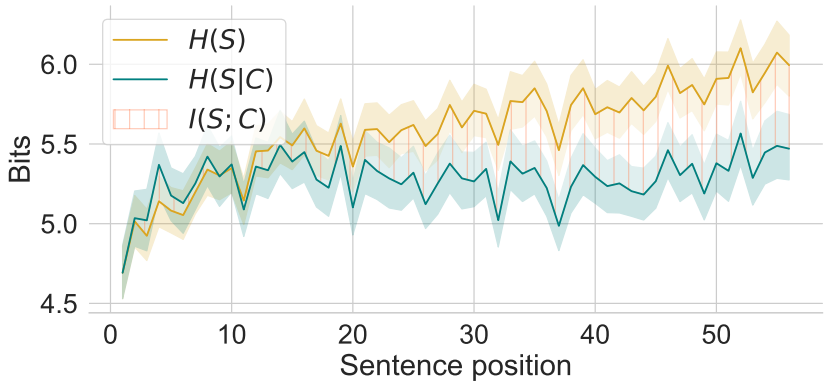
Context informativeness $I(S;C)$. Context informativeness increases with utterance position in all corpora. Its rate of increase is higher than that of $H(S)$ (recall that these two quantities must increase at a similar rate for $H(S|C)$ to remain constant). In Penn Treebank and Spoken BNC, $I(S;C)$ increases very rapidly in the initial positions; in PhotoBook, the rate of increase is more regular and yields the strongest effect in our statistical models.

Contextualised surprisal $H(S|C)$. We find no significant effect of utterance position on contextualised surprisal in Penn Treebank: $H(S|C)$ remains constant as predicted by the ERC principle. However, we observe a significant negative effect in both dialogue corpora.

Summary. The results of Experiment 1 empirically confirm Genzel and Charniak's assumption (2002) that context informativeness increases throughout discourse. $H(S)$ and $I(S;C)$, however, do not always increase together, and when they do, they grow at a different rate. In Penn Treebank, the difference in rate is sufficiently low to keep $H(S|C)$ constant but this is not the case in the dialogue corpora: in PhotoBook $I(S;C)$ increases much faster than $H(S)$, and in Spoken BNC, $H(S)$ does not increase at all. The regression coefficients are rather small but comparable to those found in prior work (Qian and Jaeger, 2011; Xu and Reitter, 2018). In sum, we find that the ERC principle holds in our corpus of written monologue, but it incorrectly predicts the rate of information in our two dialogue corpora.



(a) Penn Treebank



(b) PhotoBook

(c) Spoken BNC

Figure 7.1: Decontextualised surprisal $H(S)$, contextualised surprisal $H(S|C)$, and context informativeness $I(S;C)$ against sentence position. Bootstrapped 95% confidence intervals.

7.2.2 Experiment 2: Is surprisal uniform?

Experiment 1 suggests that constancy may not be the best descriptor for patterns of contextualised surprisal, particularly in dialogue. In Experiment 2, we test whether these patterns can be described as uniform. Collins (2014) proposes two criteria to assess uniformity: local predictability and global centrality. *Local predictability* measures whether surprisal changes in a slow and predictable way from one linguistic unit to the next, as this is expected to reduced the addressee's processing effort and the chances of miscommunication. *Global centrality* measures to what extent surprisal estimates cluster around a fixed value; this criterion is directly derived from the noisy channel model, predicting that language is transmitted at a stable rate, close to the channel capacity (Shannon, 1948). These measures were originally defined by Collins (2014) to test for uniformity of word-level surprisal within a sentence; here, we apply them at the sentence level within a discourse. Since they assess uniformity according to different criteria, it is sufficient for one of them to hold to consider information profiles uniform.

We measure global centrality and local predictability of $H(S|C)$ within each document of a corpus. In particular, we calculate *local predictability* as the mean squared difference in $H(S|C)$ between two consecutive utterances:

$$LP = -\frac{1}{N}\sum_{i=2}^{N}\left(H\left(S_i|C_i\right) - H\left(S_{i-1}|C_{i-1}\right)\right)^2 \tag{7.2}$$

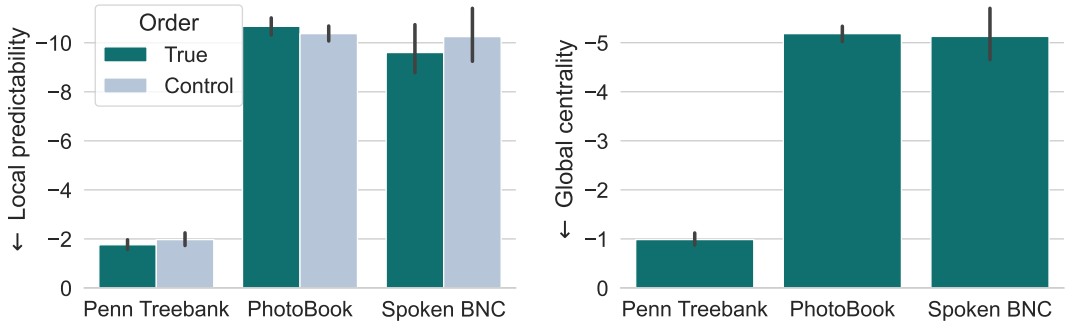
where N is the number of utterances in a document or dialogue. We also compute local predictability on 100 randomly shuffled versions of a document, and compare the true and control scores. If local uniformity of information has an influence on speakers' choices, we should find a significant difference between true and control local predictability scores. *Global centrality* is the negative variance of contextualised surprisal of all utterances in a document:

$$GC = -\frac{1}{N}\sum_{i=1}^{N}\left(H\left(S_i|C_i\right) - \mu\right)^2 \tag{7.3}$$

where μ is the mean surprisal over the utterances in a document.

Our key results are visualised in Figure 7.2. We now discuss the two measures of uniformity in turn.

Local predictability. We find the highest degree of local predictability in the Penn Treebank articles; $H(S|C)$ estimates for PhotoBook and Spoken BNC show much lower levels of uniformity according to this criterion (see Fig. 7.2a). For all three corpora, the local predictability of the true documents is not significantly different from that of shuffled documents: this suggests that, within discourse, the pressure for maintaining the levels of surprisal locally similar is not as pronounced as it is within a sentence (e.g., Jaeger and Levy, 2007; Collins, 2014; Meister et al., 2021).



(a) Local predictability (b) Global centrality

Figure 7.2: Per-document uniformity of contextualised surprisal $H(S|C)$. Bootstrapped 95% confidence intervals.

Global centrality. The written texts of the Penn Treebank exhibit a higher degree of global centrality than both written and spoken dialogues (see Fig. 7.2b). This is in line with our findings for Experiment 1: in Section 7.2.1, we reported no effect of utterance position on $H(S|C)$ in the Penn Treebank, and now we observe that all information estimates in the Penn Treebank documents tend to cluster around a fixed value. Overall, these results indicate that in the Penn Treebank newspaper articles, information is transmitted at a constant and globally uniform rate. In the dialogue corpora, where we found the rate of increase of $H(S)$ and $I(S;C)$ to be significantly different, $H(S|C)$ values are less uniformly distributed according to the global centrality criterion.

Summary. Experiment 2 shows that utterance surprisal is significantly more uniform in written monologue than in written and spoken dialogue, both at a local and at a global level. A possible explanation for this may be the fact that while in newspaper articles uniformity depends on the linguistic choices of a single writer, dialogue utterances are produced online by two speakers, which makes it harder to keep levels of surprisal locally and globally uniform. Furthermore, comparing the local predictability scores of original and shuffled documents, we find that local predictability is not an accurate descriptor of information transmission patterns in discourse.

7.3 Surprisal within contextual units

We have seen that in dialogue, where efficient strategies of information exchange need to be coordinated between two speakers, it is more difficult to observe constant or uniform information profiles. However, our first two experiments examined surprisal throughout entire conversations. Trends of constancy or uniformity

may only become visible if we zoom in on structural units that determine the type and size of the overall relevant context. Genzel and Charniak (2002, 2003), for example, show that relevant contextual cues in texts are lexical (writers tend to reuse words that have already appeared in the discourse) and topically determined, as given by the paragraph structure of texts. In dialogue, defining a topically relevant contextual unit is not straightforward. Xu and Reitter (2018), for example, use a topic segmentation algorithm to identify relevant units in open-domain dialogues and show that surprisal is influenced by topic shift.

In this second study, we exploit the inherent (task-related) structure of task-oriented dialogues to test the constancy hypothesis within contextual units of different type and size. We focus on constancy, rather than uniformity, as our previous experiments suggested this is a better descriptor of information transmission patterns.

7.3.1 Contextual units and hypotheses

We analyse two corpora of task-oriented English dialogues: Map Task (Anderson et al., 1991)[1] and PhotoBook (Haber et al., 2019)[2]. See Section 6.3 for more details; dialogue excerpts can be found in Appendix B.1.

Map Task. We consider two types of contextual unit: *a) the overall dialogue*: a series of landmarks are described in succession to help the instruction follower draw a path towards a goal location; *b) a dialogue transaction*: a dialogue excerpt related to reaching a certain landmark, manually annotated as part of the corpus. For both types of contextual unit, we also construct versions where we use the Map Task dialogue act annotation to filter out utterances exclusively consisting of backchannels and other grounding acts (*'okay'*, *'mmhmm'*) common in spoken language.[3] This results in contextual units that focus on information-transmission dialogue acts and are more referentially coherent. We hypothesise that, in Map Task, the constancy of contextualised surprisal (and the corresponding increase in decontextualised surprisal and context informativeness) will be more visible at the transaction level, where the context is more topically coherent, than at the dialogue level, where a dozen different landmarks are brought up in succession—in particular when only information-transmission dialogue acts are taken into account.

PhotoBook. We investigate the following types of contextual unit: *a) the overall dialogue*: throughout a game, all the photographs are about a certain domain

[1] `http://groups.inf.ed.ac.uk/maptask`

[2] `https://dmg-photobook.github.io`

[3] We exclude acknowledgements, attention and agreement checks, and pre-initiating moves.

Contextual unit	**All dialogue acts**			**Information-transmission acts**		
	$H(S)$	$H(S\|C)$	$I(S;C)$	$H(S)$	$H(S\|C)$	$I(S;C)$
Dialogue	-0.011*	-0.067	0.310	0.005*	-0.016*	0.085
Dialogue (givers)	-0.001*	-0.053	0.297	0.013*	-0.013*	0.101
Dialogue (followers)	-0.028*	-0.090	0.334	-0.006*	-0.014*	0.047*
Transaction	0.002*	-0.050	0.263	0.040	-0.005*	0.207
Transaction (givers)	0.015*	-0.038	0.245	0.068	0.028	0.193
Transaction (followers)	-0.012*	-0.070	0.307	-0.022*	-0.072	0.226

Table 7.2: **Map Task**. mixed effects linear regression coefficients for utterance position within the relevant contextual unit. Predicted variables: decontextualised surprisal $H(S)$, contextualised surprisal $H(S|C)$, and context informativeness $I(S;C)$. Continuous predictors are mean-centred and scaled by 2 s.d. Asterisk indicates *not* statistically significant predictors. Surprisal estimates in transactions are computed as a function of the transaction context (i.e., C=transaction).

(e.g., food or dogs); *b) a dialogue round*: different images are described in succession as participants try to figure out which ones they share in a given round; *c) an image reference chain*: the (non-adjacent) utterances that refer to a certain image across rounds (we use the automatic annotation of referring utterance chains by Takmaz et al., 2020). In PhotoBook, we expect the strongest effects at the level of reference chains. Chains are determined both topically, by the target image, and lexically, by the conceptual pacts established in previous mentions of a target (Brennan and Clark, 1996). In rounds and dialogues, where several different images are described, topic and lexical choices are constrained by the image domain but the vocabulary used in previous utterances is more varied. We thus expect the effect to be less pronounced at these two levels.

7.3.2 Experiment 3: Constancy within contextual units

We use the language models, GPT-2 fine-tuned on the 70% training set of Map Task and PhotoBook, to estimate the surprisal of the 30% held-out portion of each respective corpus, and count utterance positions from the beginning of the relevant structural unit. To test whether utterance surprisal remains uniform, we fit a linear mixed effect model using the logarithm of surprisal as response variable and the logarithm of utterance position and utterance length as predictors. We include a random slope for the utterance position and a random intercept term grouped by distinct dialogues, which allows us to model variation among individual speakers as a function of their addressee. We examine patterns of decontextualised and contextualised surprisal ($H(S)$ and $H(S|C)$) as well as context informativeness ($I(S;C)$). Tables 7.2 and 7.3 summarise the results.

Contextual unit	$H(S)$	$H(S\|C)$	$I(S;C)$
Dialogue	0.076	-0.026	0.509
Round (C=round)	-0.002*	-0.039	0.180
Round (C=dialogue)		0.027	-0.148
Chain (C=chain)	0.062	-0.137	0.812
Chain (C=dialogue)		-0.048	0.465

Table 7.3: **PhotoBook**. mixed effects linear regression coefficients for utterance position within the relevant contextual unit. Predicted variables: decontextualised surprisal $H(S)$, contextualised surprisal $H(S|C)$, and context informativeness $I(S;C)$. Continuous predictors are mean-centred and scaled by 2 s.d. Asterisk indicates *not* statistically significant predictors. C=dialogue indicates that the surprisal estimate is computed as a function of the dialogue context; C=chain indicates that surprisal is computed conditioned on the previous utterances in the reference chain.

Map Task. In line with results from Experiment 1, when we take entire Map Task dialogues as the contextual unit, we do not find a positive effect of utterance position on decontextualised surprisal and we observe a slight decrease in contextualised surprisal (see Table 7.2 and Figure 7.3). Again, this is driven by an increase in context informativeness. The same trends hold for both speaker roles: instruction givers and followers. However, focusing on information-transmission dialogue acts, we observe that utterance position has no significant effect on contextualised surprisal—i.e., that surprisal does remain constant if we filter out backchannels and other grounding acts.

The types of dialogue act considered also affect our results on transactions. We fail to find an effect in transactions with backchannels but the linear mixed effect models show a positive effect of utterance position within transactions without backchannels for decontextualised surprisal and no effect for contextualised surprisal (see also Figure 7.4). We attribute these findings to the nature of the task. Over the course of a dialogue, speakers traverse a map naming different landscape features and are thus unable to establish more than a minimal level of linguistic routine at the dialogue level. Transactions, on the other hand, correspond to more referentially constrained subtasks; this becomes more evident when information-transmission dialogue acts are isolated from transmission-coordination acts. Analysing the instruction giver and follower information-transmission utterances independently reveals that there is no significant effect of position on decontextualised surprisal for instruction followers; the overall positive effect is driven by the instruction givers (see also Figure 7.5). This reflects the asymmetric nature of information transmission in Map Task dialogues.

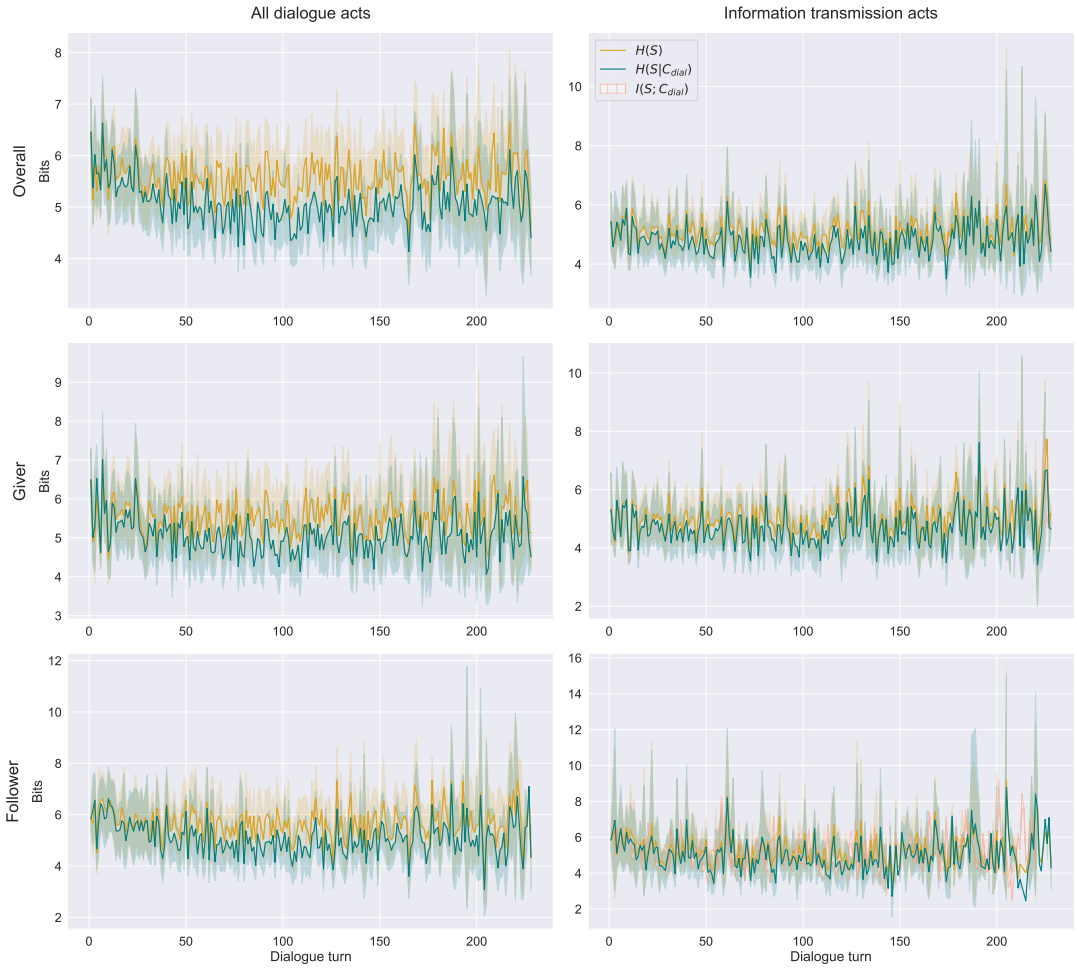


Figure 7.3: Map Task, surprisal vs. utterance position in dialogues.

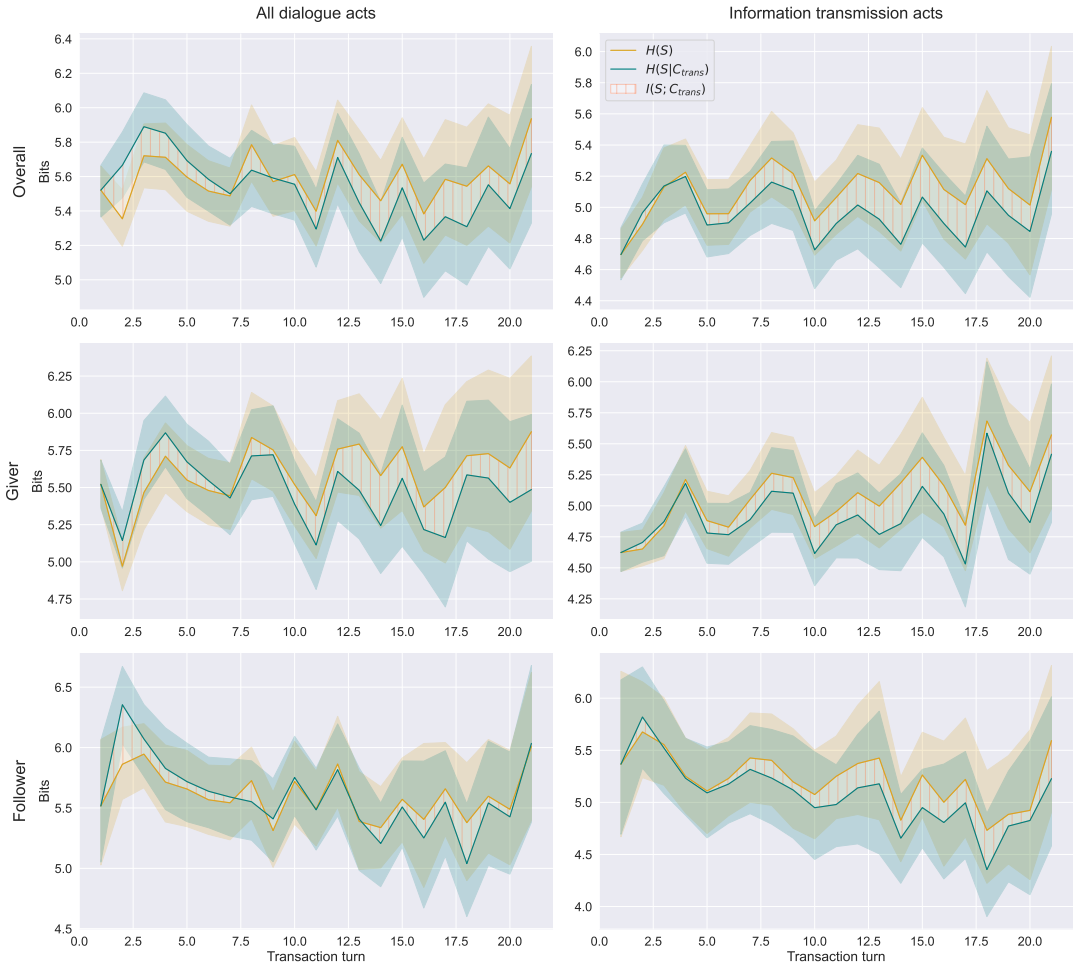


Figure 7.4: Map Task, utterance position in transactions.

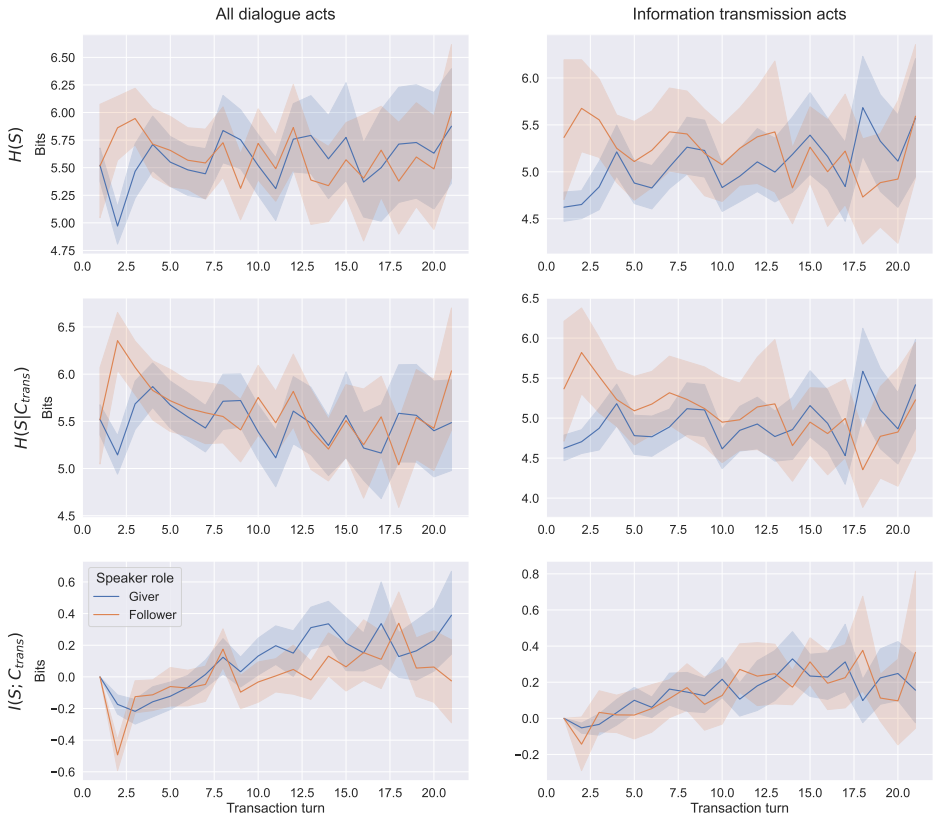


Figure 7.5: Map Task, surprisal vs. utterance position in transactions. Instruction givers vs. followers.

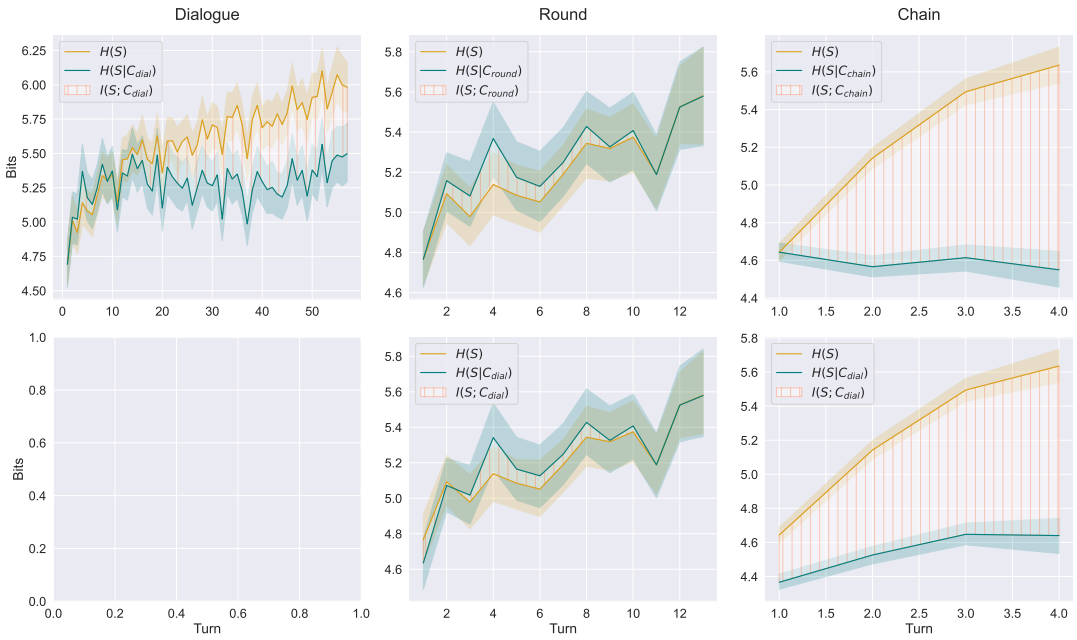


Figure 7.6: PhotoBook, surprisal vs. utterance position. All contextual units.

PhotoBook. For both rounds and reference chains, we examine contextualised surprisal estimates obtained by conditioning the language model either on the whole dialogue (up to the current utterance) or on the smaller contextual unit (round and chain respectively). The results are summarised in Table 7.3.

The effect of position on decontextualised surprisal is positive within the PhotoBook dialogues, yet contextualised surprisal (with utterance length factored out) decreases as a result of a strong increase in context informativeness over dialogue turns. Figure 7.6 shows a consistently increasing pattern for decontextualised surprisal, providing evidence that participants optimise their information-transmission strategy throughout PhotoBook games.

Within PhotoBook rounds, decontextualised surprisal does not increase, nor does contextualised surprisal remain constant. Because multiple images are discussed in a round, this contextual unit seems not to capture the relevant context of individual dialogue utterances nor be large enough to display the participants' overall information transmission strategy that we observe at the dialogue level.

Finally, as hypothesised, the effect of position is positive on decontextualised surprisal at the reference chain level. Constancy does not hold, however, regardless of whether surprisal estimates are conditioned on the whole dialogue or only on the reference chain. Increases in context informativeness are particularly large in magnitude, especially when conditioning on reference chains. As participants re-refer to an image over the game, they increase the information density of their messages (as shown in Figure 7.6) and also decrease message length (Kendall's

correlation between position in chain and length is $\tau = -0.268, p < 0.001$). Thus, as reference chains unfold, the *reduction* process observed by Takmaz et al. (2020) is complemented by *information compression*.

The relatively low magnitude of the fixed effect as well as that of the correlation between utterance length and chain position, however, suggest that the process we see at play is not only one of compression and reduction. We often observe that the fourth position in a chain comes with a decrease in decontextualised surprisal, perhaps indicating that once a conceptual pact has been established between interlocutors, referential expressions can be significantly simplified without losing referential power—as in the following reference chain (decontextualised surprisal estimates in parenthesis):

1. *'Man eating slice of pizza'* (0.69)
2. *'last one for me is guy with pizza'* (0.78)
3. *'pizza eater'* (0.91)
4. *'pizza'* (0.67)

7.4 Discussion and conclusions

In this chapter, we have examined some central tenets of the classic information-theoretic model of communication. In contrast to previous work, we have used language models to obtain surprisal estimates for utterances *within their discourse context* ($H(S|C)$), and we have measured context informativeness ($I(S;C)$) as the reduction in utterance surprisal contributed by discourse with respect to out-of-context estimates ($H(S)$). This has allowed us to directly model the information transmission profiles of written texts as well as written and spoken dialogues and, thereby, to test whether they follow the rational communicative strategies predicted by prominent hypotheses of rational use of the communication channel.

Experiments 1 and 2. In our first two experiments, we have found that in American English newspaper articles, $H(S|C)$ remains stable as predicted by the theory. This is not the case, however, for spoken British English open-domain dialogues, nor for written English task-oriented dialogues: here, $H(S|C)$ decreases, albeit moderately, as the utterance position grows. We suggest that this is the result of the uneven rates of increase measured for $H(S)$ and $I(S;C)$—the latter increases faster than the former in all corpora under examination. We find the strongest $I(S;C)$ increase in the PhotoBook dialogues, where topic is determined by a game's image domain and, by task design, participants produce multiple subsequent utterances to describe the same images over game rounds. Correct interpretation of subsequent references (McDonald, 1978) requires indeed access to the shared knowledge accumulated by speakers during dialogue. We observe the second strongest $I(S;C)$ increase in the Penn Treebank articles, where topic

is consistent throughout the text but new information keeps being conveyed from the beginning to the end of the discourse. The weakest increase takes place in the Spoken BNC: topic is more likely to change during the course of an open-domain dialogue and, with topic shifts, the previously established common ground becomes less relevant for the prediction of new linguistic material.

The lower rates of increase of $H(S)$, on the other hand, can be due to the limits imposed on lexical choice by grammar and style. In PhotoBook, where participants write freely in a chat interface, the increase is stronger than in the more formal newspaper articles of the Penn Treebank. However, the stable $H(S)$ trends in the Spoken BNC suggest that this is only one side of the coin. The theory predicts that when context is more informative, speakers will increase the density of their sentences to be more efficient. But speakers do not need to be always efficient in open-domain conversations, where the pure information transmission goal is perhaps overweighted by other goals, such as social goals, which are not taken into account by the theory.

Another empirical finding that is not in line with expectations derived from the theoretical framework is that uniformity of surprisal across consecutive utterances (local predictability) is not a good predictor of the information transmission profiles of the texts and dialogues we analysed. Local uniformity may be more relevant for lower-level linguistic signals as they come in a much faster succession: speakers want to avoid sudden changes in surprisal to reduce comprehension effort; yet, at the discourse level, changes in surprisal are less abrupt as they are spread throughout an entire utterance, thus giving the addressee time to adapt gradually to the higher surprisal of the larger transmission unit. Global centrality seems to be a more faithful criterion of uniformity, in particular for the articles of the Penn Treebank. In other words, utterances are not so much produced to limit the difference in surprisal with respect to the previous sentence, but rather to maintain the overall transmission rate stable in the articles. PhotoBook and Spoken BNC show a significantly lower degree of uniformity than the Penn Treebank, measured both as local predictability and global centrality: in dialogue, an efficient strategy of information exchange needs to be coordinated between two speakers, which can make it more difficult to obtain uniform information profiles.

Experiment 3. We concluded from our first two experiments that the classical model of communication may be too simplistic for discourse, where the units of information are more complex. A first issue has to do with identifying the relevant contextual components, which are determined, at least, by the internal structure of the discourse (Genzel and Charniak, 2003) and by topic shifts (Qian and Jaeger, 2011; Xu and Reitter, 2018). In our third experiment, we have related the properties of task-determined contextual units to patterns of information transmission and have hypothesised that the UID principle holds to a stronger degree in more topically coherent and reference-specific contextual units.

Our hypotheses are confirmed in PhotoBook, where we find evidence that dialogue participants use rational strategies of information transmission over an entire dialogue. We do not observe uniformity of information in the Map Task dialogues and transactions as a whole, similarly to other negative results in interactive settings (e.g., Vega and Ward, 2009; Doyle and Frank, 2015b). Yet the effect is present within Map Task transactions when we restrict our analysis to information-transmission dialogue acts: these make for a more topically and referentially coherent contextual unit. Indeed, the organisation of context can be complex in dialogues. We have shown that theoretically motivated contextual units such as reference chains in PhotoBook and information-transmission acts in Map Task transactions are good candidates to characterise the relevant context over which participants deploy strategies of information compression.

Limitations. While the classic information-theoretic framework of communication adopted in this study assumes a single addressee across documents, communication is shaped by the identity and the characteristics of multiple addressees (Brennan and Clark, 1996; Brown-Schmidt et al., 2015).

Moreover, our estimates of surprisal assume a static addressee (the language model is only fine-tuned once on the corpus, but never updated within a single dialogue) whereas true addressees adapt on-the-fly: e.g., van Schijndel and Linzen (2018) show that endowing a language model with a simple adaptation mechanism improves predictions of human reading times compared to a non-adaptive model. We will use this kind of model in the next chapter.

Finally, the framework condenses production and comprehension effort in a single estimate. Future work should study strategies of information transmission in discourse using a model of communication, such as the Rational Speech Act model (Frank and Goodman, 2012a), that includes production costs more explicitly and that allows accompanying cognitive costs with social costs—e.g., those related to the goal of the linguistic interaction. Zaslavsky et al. (2021) recently showed that the RSA model optimises the trade-off between expected utility and communicative effort, and that it is directly related to Rate-Distortion theory (Shannon, 1948)—the branch of information theory that formalises the effect of limited transmission resources on communicative success.

Outlook. Overall, the studies presented in this chapter provide new empirical evidence on language production in dialogue, which we believe can inform the development of natural language generation models. Our findings suggest that models that take relevant contextual units into account (Takmaz et al., 2020; Hawkins et al., 2020a) are better suited for reproducing human patterns of information transmission, and confirm that training objectives and sampling strategies that enforce a uniform organisation of information density (Meister et al., 2020; Wei et al., 2021) are a promising avenue for language models.

Chapter 8

The facilitating effect of construction repetition

The content of this chapter is based on the following publication:

> Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2022. Construction Repetition Reduces Information Rate in Dialogue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), pages 665–682, Online only. Association for Computational Linguistics.

Mario and Arabella jointly produced the research idea. Mario performed the experiments with Arabella and Raquel's supervision. Mario wrote the article, Arabella and Raquel reviewed and contributed to the writing. The text in this chapter overlaps with that of the original publication.

8.1 Introduction

The repeated use of particular configurations of structures and lexemes, *constructions*, is pervasive in conversational language use (Tomasello, 2003; Goldberg, 2006). Such repetition can be understood as surface level evidence of processes of coordination (Sinclair and Fernández, 2021) or 'interpersonal synergy' between conversational partners (Fusaroli et al., 2014). Speakers may use repetitions to successfully maintain common ground with their interlocutors (Brennan and Clark, 1996; Pickering and Garrod, 2004), because they are primed by their recent linguistic experience (Bock, 1986), or to avoid a costly on-the-fly search for alternative phrasings (Kuiper, 1995). At the same time, repetitions are also advantageous for comprehenders. Repeating a sequence of words positively reshapes expectations for those words, allowing comprehenders to process them more rapidly (for a review, see Bigand et al., 2005).

As speakers are known to take into consideration both their own production cost and their addressee's processing effort (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989; Frank and Goodman, 2012a), its two-sided processing advantage, as described above, makes construction repetition an efficient, cost-reducing communication strategy. In this study, we investigate whether and how information processing properties of repetitions shape patterns of surprisal in open-domain spoken dialogue, trying to reconcile principles of efficiency and rationality with our findings on dialogue from the previous chapter.

As we have seen in Chapter 7, the constancy hypothesis has found empirical support for written language production (Genzel and Charniak, 2002, 2003; Qian and Jaeger, 2011, as well as our experiments) but results on dialogue are mixed (Vega and Ward, 2009; Doyle and Frank, 2015b,a; Xu and Reitter, 2018, and our experiments), with some studies suggesting a decreasing information rate over the course of dialogues (Vega and Ward, 2009, and again, our experiments in the pervious chapter). We hypothesise that the decreasing surprisal trends observed in dialogue may be associated with construction repetition. We conjecture that speakers use construction repetition as a strategy for information rate mitigation, by padding the more surprising, or information dense, parts of their utterances with progressively less information dense constructions—leading to an overall decrease in information rate over the course of a dialogue.

We extract occurrences of fully lexicalised constructions (see Table 8.1 for examples) from a corpus of open-domain spoken dialogues and use a Transformer-based neural language model to estimate their contribution to utterance surprisal. First, we confirm that constructions indeed exhibit lower surprisal than other expressions and that surprisal further decreases when constructions are repeated. Then, we show that the decreasing trend of surprisal observed *over utterances*—which contradicts the Entropy Rate Constancy principle—is driven by the increasing mitigating effect of construction repetition, measured as a construction's (increasingly) negative contribution to the surprisal of its containing utterance,

SPXV	**SAXQ**	**S9YG**
want to be with him	*it on the television*	I bet you can
shit like that	*for a family*	yeah I used to
I can be	think that's a	*go to bed*
to see her	*the orient express*	and I love
and she just	one thing that	*the window and*
I quite like	*one of my favourites*	and I think it's
you don't like	*on the television*	yeah I think so
and you're like	yes yeah I	*the same people*
going to go	erm I think	is she in
you're going to	a really good	*lock the door*

Table 8.1: Top 10 constructions from three dialogues of the Spoken BNC, sorted according to the PMI between a construction and its dialogue (Section 8.6.1). Referential constructions in italics (Section 8.3). Headers correspond to the dialogues' IDs in the corpus.

what we call its *facilitating effect*.

In sum, our study provides new empirical evidence that dialogue partners use construction repetition as a strategy for information rate mitigation, which can explain why the rate of information transmission in dialogue, in contrast to the constancy predicted by the theory (Genzel and Charniak, 2002), is often found to decrease.

8.2 Constructions

This work focuses on *constructions*, seen as particular configurations of structures and lexemes in usage-based accounts of natural language (Tomasello, 2003; Bybee, 2006, 2010; Goldberg, 2006). According to these accounts, models of language processing must consider not only individual lexical elements according to their syntactic roles but also more complex form-function units, which can break regular phrasal structures—e.g., *'I know I'*, *'something out of'*. We further focus on fully lexicalised constructions (sometimes called *formulaic expressions*, or *multi-word expressions*). Commonly studied types of constructions are idioms (*'break the ice'*), collocations (*'pay attention to'*), phrasal verbs (*'make up'*), and lexical bundles (*'a lot of the'*). In Section 8.3, we explain how the notion of lexicalised construction is operationalised in the current study; Table 8.1 shows examples from three dialogues in the Spoken BNC corpus (Love et al., 2017).

A common property of constructions is their frequent occurrence in natural language. As such, they possess what, in usage-based accounts, is sometimes

referred to as 'processing advantage' (Conklin and Schmitt, 2012; Carrol and Conklin, 2020). Evidence for the processing advantage of construction usage has been found in reading (Arnon and Snider, 2010; Tremblay et al., 2011), naming latency (Bannard and Matthews, 2008; Janssen and Barber, 2012), eye-tracking (Underwood et al., 2004; Siyanova-Chanturia et al., 2011), and electrophysiology (Tremblay and Baayen, 2010; Siyanova-Chanturia et al., 2017). In this study, we model this processing advantage as reduced surprisal and show that it can mitigate information rate throughout entire dialogues.

8.2.1 What makes a construction?

Constructions can be classified according to multiple criteria (Titone and Connine, 1994; Wray, 2002; Columbus, 2013). This section presents common criteria and discusses why our study is agnostic with respect to most of them.

Compositionality. This criterion is typically used to separate idioms from other formulaic expressions, although it is sometimes referred to as *transparency* to underline its graded, rather than binary, nature. There is no evidence, however, that the processing advantage of idioms differs from that of compositional phrases (Tabossi et al., 2009; Jolsvai et al., 2013; Carrol and Conklin, 2020). Therefore, we ignore this criterion in the current study.

Literal plausibility. This is used as a criterion to discriminate among different types of idioms (Titone and Connine, 1994; Titone and Libben, 2014)—as compositional phrases are literally plausible by definition. Because we ignore distinctions made on the basis of compositionality, we do not use this criterion.

Meaningfulness. Meaningful expressions are idioms and compositional phrases (e.g., *'on my mind'*, *'had a dream'*) whereas sentence fragments that break constituency boundaries (e.g., *'of a heavy'*, *'by the postal'*) are considered less meaningful (as measured in norming studies, e.g., by Jolsvai et al., 2013). There is some evidence that the meaningfulness of multi-word expressions correlates with their processing advantage (Jolsvai et al., 2013); yet if expressions are particularly frequent, they present processing advantages in spite of breaking regular phrasal structures (Bybee and Scheibman, 1999; Tremblay et al., 2011). Moreover, utterances that break regular constituency rules are particularly frequent in spoken dialogue data (e.g., *'if you could search for job and that's not'*, *'you don't wanna damage your relationship with'*). For these reasons, we do not exclude constructions that span multiple constituents from our analysis.

Schematicity. This criterion distinguishes expressions where all the lexical elements are fixed from expressions with 'slots' that can be filled by varying lexical

elements. In this study, we focus on fully lexicalised constructions.

Familiarity. This is a subjective criterion that strongly correlates with objective frequency (Carrol and Conklin, 2020). Human experiments would be required to obtain familiarity norms for our target data, and the resulting norms would only be an approximation of the familiarity judgements of the actual speakers we analyse the language of. Therefore, we ignore this criterion in the current study.

Communicative function. Formulaic expressions can fulfil a variety of discourse and communicative functions. Biber et al. (2004), e.g., distinguish between stance expressions (attitude, certainty with respect to a proposition), discourse organisers (connecting prior and forthcoming discourse), and referential expressions; and for each of these three primary discourse functions, more specific subcategories are defined. This type of classification is typically done a posteriori—i.e., after a manual analysis of the expressions retrieved from a corpus according to other criteria (Biber and Barbieri, 2007). In the BNC, for example, we find epistemic lexical bundles (*'I don't know'*, *'I don't think'*), desire bundles (*'do you want to'*, *'I don't want to'*), obligation/directive bundles (*'you don't have to'*), and intention/prediction bundles (*'I'm going to'*, *'it's gonna be'*). In this study, we distinguish between referential and non-referential constructions.

8.3 Data: Extracting repeated constructions

We conduct our study on the Spoken BNC corpus, presented in Section 6.3. We focus on the 622 dialogues that feature only two speakers, and randomly split them into a 70% fine-tuning set (to be used as described in Section 8.4) and a 30% analysis set (used in our experiments, as described in Section 8.5 and Section 8.6). Table 8.2 shows some basic statistics of the dialogues used in this study.

We define constructions as multi-word sequences repeated within a dialogue. To extract constructions from each dialogue, we use the sequential pattern mining method proposed by Duplessis et al. (2017a,b, 2021), which treats the extraction task as an instance of the longest common subsequence problem (Hirschberg, 1977; Bergroth et al., 2000).[1] We modify it to not discard multiple repetitions of a construction that occur in the same utterance. We focus on constructions of at least three tokens, uttered at least three times in a dialogue by any of the dialogue participants. Repeated sequences that appear less than twice outside of a larger repeated construction in a given dialogue (e.g., *'think of it'* vs. *'think of it like'*) are discarded. We also exclude sequences containing punctuation marks or which consist of more than 50% filled pauses (e.g., *'mm'*, *'erm'*).[2]

[1] Their code is freely available at `https://github.com/GuillaumeDD/dialign`.

[2] The full list of filled pauses can be found in Appendix B.5.

	Mean ± Sd	Median	Min	Max
Dialogue length (# utterances)	736 ± 599	541.5	67	4,859
Dialogue length (# words)	7,753 ± 5,596	6,102	819	39,575
Utterance length (# words)	11 ± 15	6	1	982

Table 8.2: Two-speaker dialogue statistics, Spoken BNC.

Applying the described extraction procedure to the 187 dialogues in the analysis split of the Spoken BNC yields a total of 5,893 unique constructions and 60,494 occurrences. Further statistics of the extracted constructions are shown in Table 8.3.

For analysis purposes, we distinguish between referential and non-referential constructions. We label a construction as *referential* if it includes nouns, unless the nouns are highly generic.[3] Referential constructions are mostly topic-determined; examples are *'playing table tennis'*, *'a woolly jumper'*, *'a room with a view'*. The remaining constructions are labelled as *non-referential*. These mainly include topic-independent expressions and conversational markers, such *'a lot of'*, *'I don't know'*, and *'yes of course'*. Our dataset consists of 5,291 referential and 55,203 non-referential construction occurrences, corresponding to 1,143 and 4,750 construction forms. Table 8.1 includes examples of both construction types.

	Mean ± Sd	Median	Max
Construction Length	3.27 ± 0.58	3	7
Construction Frequency	4.29 ± 3.04	3	70
Constructions per Dialogue	325.34 ± 458.64	149	2,817
Referential	30.96 ± 39.75	19	346
Non-Referential	296.88 ± 424.17	134.5	2,530
Utterance Length	31.19 ± 36.19	21	959

Table 8.3: Construction statistics for our analysis split of the Spoken BNC. *Construction Length*: number of words in a construction. *Construction Frequency*: occurrences of a given construction in a dialogue. *Constructions per Dialogue*: occurrences of all constructions in a dialogue. *Utterance Length*: number of words in utterances containing a construction. The minimum is always 3 by design (Section 8.3). The difference between referential and non-referential is only significant for *Constructions per Dialogue*, so we report separate statistics for this variable.

[3]We define a limited specific vocabulary of generic nouns (e.g., *'thing'*, *'fact'*, *'time'*); full vocabulary in Appendix B.5.

8.4 Experimental setup

In this section, we define our proposed measure of facilitating effect and present the adaptive language model used to produce surprisal estimates.[4]

8.4.1 Measures of surprisal

In Section 6.2.1, we defined the surprisal of a word choice w_i as the negative logarithm of the corresponding word probability, conditioned on the utterance context $S_{:w_i}$ (i.e., the words that precede w_i in utterance S) and on the relevant discourse context C (Equation 6.4). We define the discourse context C as the local dialogue context, i.e., as the 50 tokens that precede the first word in the utterance.[5] We use tokens as a unit of context size, rather than utterances, since they more closely correspond to the temporal units used in previous work (e.g., Reitter et al., 2006), and since the length of utterances can vary significantly (see Table 8.2). To measure the surprisal of a construction c, we average over word-level surprisal values:

$$H(c; S_{:c}, C) = \frac{1}{|c|} \sum_{w_i \in c} H(w_i | S_{:c}, C) \tag{8.1}$$

The above surprisal formulation can be applied constructions and entire utterances, but it does not qualify the relationship between the two. We also measure the surprisal change (increase or reduction in information rate) contributed by a construction c to its containing utterance, which we call the *facilitating effect* of a construction. Facilitating effect is defined as the logarithm of the ratio between the surprisal of a construction and that of its utterance context:

$$\mathrm{FE}(c; S, C) = \log_2 \frac{\frac{1}{|S|-|c|} \sum_{c \not\ni w_j \in S} H(w_j | S_{:w_i}, C)}{\frac{1}{|c|} \sum_{w_i \in c} H(w_i | S_{:c}, C)} \tag{8.2}$$

By definition, this quantity is positive when the construction has lower surprisal than its context, and negative when it has higher surprisal. When the utterance consists of a single construction, facilitating effect is set to 0.

We can expect the values produced by our surprisal and facilitating effect measurements (Equations 8.1 and 8.2, respectively) to correlate: it is more likely for a construction to have a (positive) facilitating effect if its surprisal is low. When a construction's surprisal is high, the surprisal of its utterance context must

[4]Code and statistical analysis are available at `https://github.com/dmg-illc/uid-dialogue`.

[5]Building on prior work (Reitter et al., 2006) that uses a window of 15 seconds of spoken dialogue as the locus of local repetition effects, we compute the average speech rate in the Spoken BNC (3.16 tokens/second) and multiply it by 15; we then round up the result (47.4) to 50 tokens.

be even greater for facilitating effect to occur. Nevertheless, perfect correlation does not follow a priori from the definition of the two measures; we will show this empirically in Section 8.5.4.

8.4.2 Adaptive language model

To estimate the per-word conditional probabilities that are necessary to compute surprisal, we use an adaptive language model. The model is conditioned on local contextual cues via an attention mechanism (Vaswani et al., 2017) and it learns continually (see, e.g., Krause et al., 2018) from exposure to the global dialogue context. We use GPT-2 (Radford et al., 2019), a pre-trained autoregressive Transformer language model. We rely on HuggingFace's implementation of GPT-2 with default tokenizers and parameters (Wolf et al., 2020) and fine-tune the pre-trained model on a 70% training split of the Spoken BNC to adapt it to the idiosyncrasies of spoken dialogue data.[6] We refer to this fine-tuned version as the *frozen* model. We use an attention window of length $|S_{:w_i}| + 50$, i.e., the sum of the utterance length up to word w_i and the size of the local dialogue context.

As a continual learning mechanism, we use back-propagation on the cross-entropy next word prediction error, a simple yet effective adaptation approach. Following van Schijndel and Linzen (2018), when estimating surprisal for a dialogue, we begin by processing the first utterance using the frozen language model and then gradually update the model parameters after each turn. For these updates to have the desired effect, the learning rate should be appropriately tuned. It should be sufficiently high for the language model to adapt during a single dialogue, yet an excessively high learning rate can cause the language model to lose its ability to generalise across dialogues. To find the appropriate rate, we randomly select 18 dialogues from the analysis split of the Spoken BNC[7] and run an 18-fold cross-validation for a set of six candidate learning rates: 1×10^{-5}, 1×10^{-4}, ..., 1. We fine-tune the model on each dialogue using one of these learning rates and compute perplexity reduction (i) on the dialogue itself (*adaptation*) as well as (ii) on the remaining 17 dialogues (*generalisation*). We select the learning rate yielding the best adaptation over cross-validation folds (1×10^{-3}), while still improving the model's generalisation ability. See Appendix B.6.2 for further details.

8.5 Preliminary experiments

In this section, we present preliminary experiments on the surprisal of utterances and constructions, which set the stage for our analysis of the facilitating effect of construction repetition.

[6] More details on fine-tuning can be found in Appendix B.6.1.

[7] This amounts to ca. 10% of the analysis split. We use the analysis split because there is no risk of "overfitting" with respect to our main analyses.

8.5.1 Utterance surprisal

Our experiments are motivated by the mixed results on the dynamics of information rate in dialogue discussed in Chapter 7. We thus begin by testing if the Entropy Rate Constancy (ERC) principle holds in the Spoken BNC, i.e., whether utterance surprisal remains stable over the course of a dialogue. Similarly to the previous chapter, and following a procedure established in prior work (Xu and Reitter, 2018), we fit a linear mixed effect model with the logarithm of utterance position and construction length as fixed effects (we will refer to their coefficients as β), and include multi-level random effects grouped by dialogue. For the ERC principle to hold, the position of an utterance within a dialogue should have no effect on its surprisal.

Instead, we find that utterance surprisal decreases significantly over time ($\beta = -0.119, p < 0.005$, *95% c.i.* $-0.130 : -0.108$), in line with previous negative results on open-domain and task-oriented dialogue, as described in the previous chapter. The strongest drop occurs in the first ten dialogue utterances ($\beta = -0.886, p < 0.005$, *95% c.i.* $-0.954 : -0.818$) but the decrease is still significant for later utterances ($\beta = -0.043, p < 0.005$, *95% c.i.* $-0.054 : -0.032$).

8.5.2 Construction surprisal

Our hypothesis that construction repetition progressively reduces the information rate of utterances is motivated by the fact that constructions are known to have a processing advantage (see Section 8.2). This property makes them an efficient production strategy, i.e., one that reduces the speaker's and addressee's collaborative effort. Before investigating if the hypothesised information rate mitigation strategy is at play, we test whether our information-theoretic measures and the language model used to generate estimates thereof are able to capture processing advantage. We expect our framework to yield lower surprisal estimates (Equation 8.1) for constructions than for other word sequences.

Indeed, the surprisal of constructions is significantly lower than that of non-construction sequences ($t = -168.82, p < 0.005$, *95% c.i.* $-2.033 : -1.987$).[8] Constructions' surprisal is on average 2 bits lower than that of non-constructions. We conclude that our estimates of surprisal are a sensible model of the processing advantage of constructions.

[8] We extract all 3- to 7-grams from our analysis split of the Spoken BNC, excluding all n-grams that are equal to extracted constructions. We then sample, for each length n from 3 to 7, s_n non-construction sequence occurrences—where s_n is the number of occurrences of n-tokens-long extracted constructions.. The length distributions should match because length has an effect on S and FE (see Section 8.6.3).

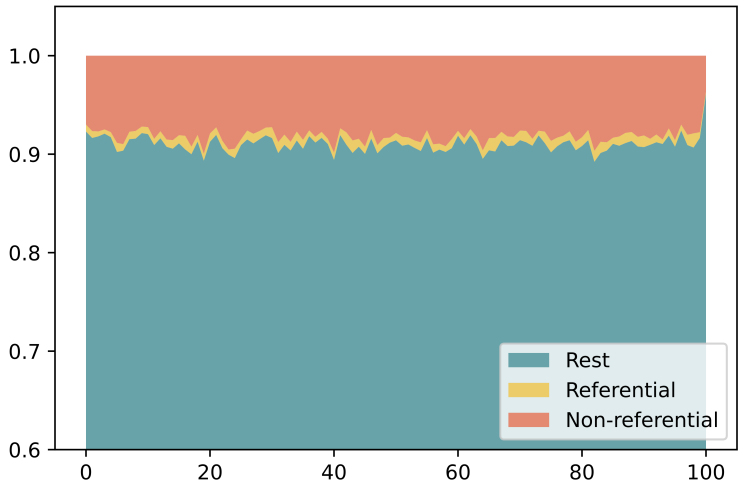


Figure 8.1: Proportion of tokens in an utterance that belong to referential constructions, non-referential constructions, and to non-construction sequences. The x axis shows percentages indicating utterance positions in the dialogue relative to the dialogue length.

8.5.3 Stable rate of construction usage

In Section 8.5.2, we confirmed that constructions have lower surprisal than other utterance material. A simple strategy to decrease utterance surprisal over dialogues (we do observe this decrease in the Spoken BNC, as described in Section 8.5.1) could then simply be to increase the rate of construction usage. To test if this strategy is at play, we fit a linear mixed effect model with utterance position as the predictor and the proportion of construction tokens in an utterance as the response variable. Over the course of a dialogue, the increase in the proportion of an utterance's tokens which belong to a construction is negligible ($\beta = 0.004, p < 0.05$, *95% c.i.* $0.001 : 0.008$). Speakers produce constructions at a stable rate (see also Figure 8.1), indicating that an alternative strategy for information rate reduction is at work.

8.5.4 Surprisal vs. facilitating effect

The facilitating effect *FE* of a construction is a function of its surprisal and the surprisal of its containing utterance (Equation 8.2). To ensure that our estimates of *FE* are not entirely determined by construction surprisal (cf. Section 8.4.1), we inspect the relation between the two measures empirically, by looking at the values they take in our dataset of constructions. We find that the Kendall's rank correlation between *FE* and surprisal is -0.623 ($p < 0.005$): although this is a

rather strong negative correlation, the fact that the score is not closer to -1 indicates that there is a substantial amount of cases where the two values are both either high or low. We indeed find examples of constructions with high surprisal H and high facilitating effect FE:

A: we'll level that right press p purchase and
B: right
A: **go back to** recommended ***(H = 5.30 FE = 1.65)***

as well cases where surprisal is low and facilitating effect is low or negative:

A: right let's go and have a drink
B yeah
A: **let's go and have** a drink ***(H = 2.10 FE = −2.21)***

These examples have been selected among occurrences with H/FE higher or lower than the mean $H/FE \pm$ sd, respectively 3.62 ± 1.48 and 0.62 ± 0.73. Further analysis shows that this is not only true for individual instances but for entire groups of constructions. In particular, although their surprisal is overall higher ($t = 13.511, p < 0.005$, *95% c.i.* $0.371 : 0.497$), referential constructions also have higher facilitating effect than non-referential ones ($t = 3.115, p < 0.005$, *95% c.i.* $0.016 : 0.072$). We conclude that the two measures capture different aspects of a construction's information profile, with facilitating effect being sensitive to both construction and utterance surprisal.

8.6 The facilitating effect of construction repetition

We now test whether constructions have a positive facilitating effect, i.e., whether they reduce the surprisal of their containing utterances. We present our main statistical model in Section 8.6.1, describe the effects of FE predictors specific to unique construction mentions in Section 8.6.2, and analyse differences between types of constructions in Section 8.6.3.

8.6.1 Method

To understand what shapes a construction's facilitating effect, we collect several motivated features that can be expected to be informative FE predictors. We fit a linear mixed effect (LME) model using (i) these features as fixed effects, (ii) FE as the response variable, (ii) and multi-level random effects grouped by dialogue and individual speaker ID. The first predictor is *utterance position*, i.e., the index of the utterance within the dialogue, which allows us to test if FE increases over the course of a dialogue. We then include predictors that distinguish different

types of repetition. Since we expect a construction mention to increase expectation for subsequent occurrences—thus reshaping their surprisal—we consider its *repetition index*, i.e., how often the construction has been repeated so far in the dialogue. Expectation is also shaped by intervening material, so we additionally track *distance*, the number of tokens separating a construction mention from the preceding one. As *FE* is the interplay between a construction and its utterance context, it is important to know whether the utterance context contains other mentions of the construction. We use a binary indicator (*previous same utterance*) to single out occurrences whose previous mention is in the same utterance; for these cases, we also count the number of same-utterance previous mentions (*repetition index in utterance*). To explore whether *FE* varies across types of expressions, we also include a binary feature indicating whether the construction is *referential* or non-referential (Section 8.3). Finally, we keep track of *construction length*, the number of tokens that constitutes a construction, and *PMI*, the pointwise mutual information between a construction and its dialogue, which is essentially a measure of the construction's frequency in the current dialogue as a function of its overall frequency in the corpus, indicating the construction's degree of interaction-specificity.[9]

To determine the fixed effects of the final model, we start with all the predictors listed above (the non-binary ones are log-transformed) and perform backward stepwise selection, iteratively removing the predictor with the lowest significance and keeping only those with $p<0.05$. All predictors make it into our final model, the one which best fits the data according to both the Akaike and the Bayesian Information Criterion. The full specification of the best model, with model fit statistics as well as fixed and random effect coefficients, can be found in Appendix B.7. The next two sections present our main findings; we report fixed effect coefficients (β), p-values (p), and 95% confidence intervals (*c.i.*).

8.6.2 Construction mentions

Our first observation is that construction usage reduces *utterance* surprisal. More precisely, we find that **facilitating effect is higher for constructions than for non-construction sequences** ($t=118.79, p<0.005$, *95% c.i.* $0.536:0.554$). Constructions have on average 62% lower surprisal than their utterance context; the average percentage drops to 7% for non-construction sequences.[10] Figure 8.2a shows the two distributions. We also observe a positive effect of utterance position on *FE* ($\beta=0.046, p<0.005$, *95% c.i.* $0.026:0.06$); that is, **the facilitating effect of constructions increases over the course of dialogues**. While the proportion of construction tokens remains stable (Section 8.5.3), their mitigat-

[9]The probabilities for the PMI calculation are obtained using maximum likelihood estimation over our analysis split of the Spoken BNC.

[10]These are the same sampled non-construction sequences as in Section 8.5.2. Their average *FE* is 0.07 ± 0.80.

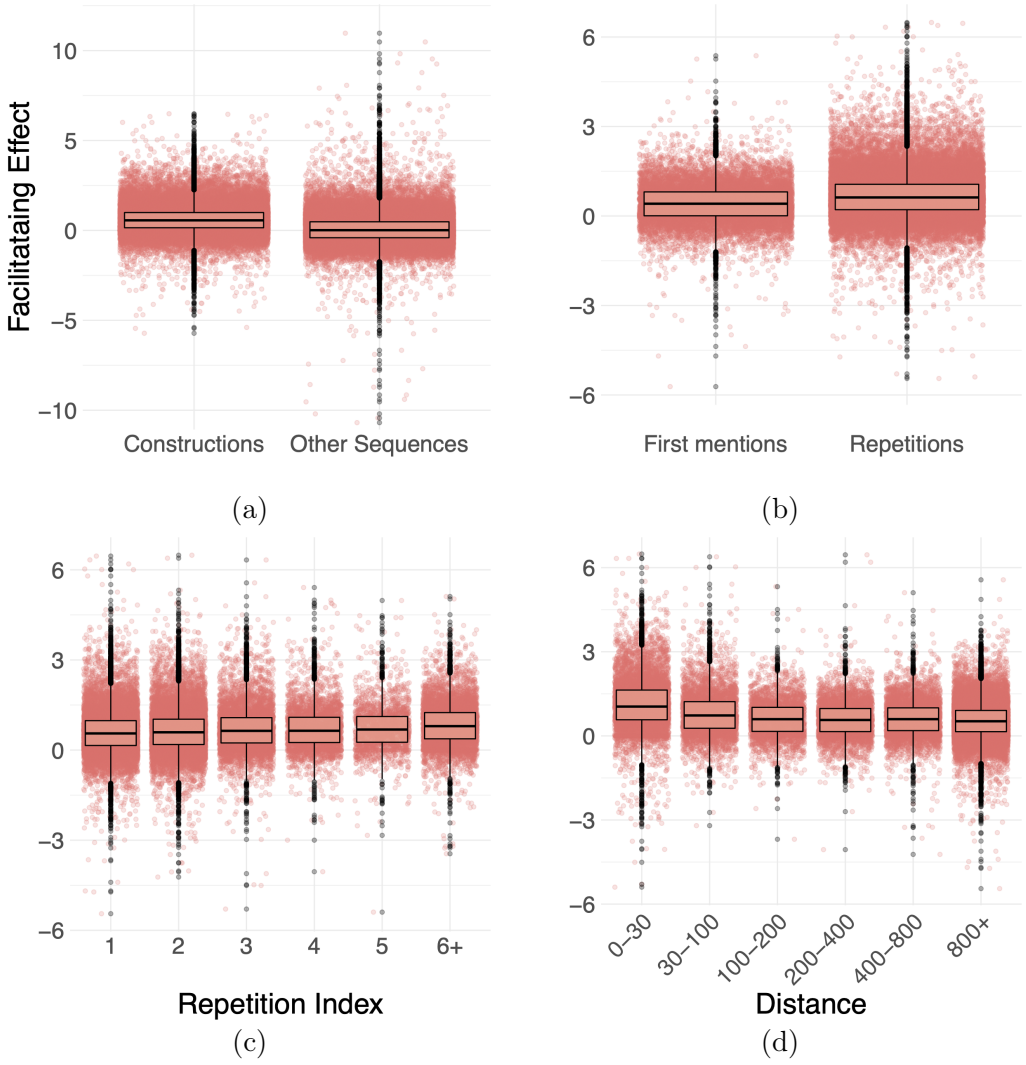


Figure 8.2: The facilitating effect (*FE*) of constructions vs. non-construction sequences (a) and of first construction mentions vs. repetitions (b); as well as *FE* vs. repetition index (c) and *FE* vs. distance from previous mention (number of words). The first distance bin (0-30) corresponds to the mean length of a turn containing a construction (Table 8.3).

Sp	RI	RI Utt	Dist	Utterance	$H(S)$	$H(c)$	$FE(c;S)$
A	0	0	-	Drink? that was what he did yeah just just to just to know that I he **might not be** a complete twat but just a fyi	5.99	4.73	0.40
B	1	0	1586	Especially for my birthday mind you I **might not be** here for	5.04	4.01	0.53
	2	1	14	mine and I went what do you mean you **might not be** here?		2.70	0.90

Table 8.4: Repetition chain for the construction *'might not be'* in dialogue SXWH of the Spoken BNC, annotated with speaker identifier (Sp), repetition index (RI), RI in utterance (RI Utt), and distance from previous mention (Dist; in tokens). $H(u)$ is the utterance surprisal, $H(c)$ and $FE(c;u)$ are the construction's surprisal and facilitating effect.

ing contribution to utterance surprisal increases throughout dialogues—perhaps since speakers are more likely to *repeat* established constructions as the dialogue develops. We indeed find that **repeated constructions have stronger facilitating effect**: there is a significant difference between the *FE* of first mentions and repetitions ($t = -38.904, p < 0.005$, *95% c.i.* $-0.265 : -0.239$), as shown in Figure 8.2b. The surprisal of repetitions is on average 68% lower than that of their utterance context; for first mentions, it is on average 42% lower.

Cumulativity and recency effects. Having observed that the mitigating contribution of constructions to utterance surprisal indeed increases with construction repetition, we now look at how the *FE* of repetitions varies as a function of their distribution over time. On the one hand, we find that **facilitating effect is cumulative**: repeating a construction reduces utterance surprisal more strongly as more mentions of the construction accumulate in the dialogue (Figure 8.2c). The effect of repetition index (i.e., how often the construction has been repeated so far in the dialogue) is positive on *FE* ($\beta = 0.079, p < 0.005$, *95% c.i.* $0.063 : 0.094$). On the other hand, the distance of a repetition from the previous mention has a negative effect on *FE* ($\beta = -0.311, p < 0.005$, *95% c.i.* $-0.328 : -0.293$). That is, **facilitating effect decays as a function of the distance between subsequent mentions**. As shown in Figure 8.2d, this is a fast decay effect: the most substantial drop occurs for low distance values. The large magnitude of this coefficient indicates that recency is an important factor for constructions to have a strong facilitating effect. Indeed, almost one third (31.8%) of all repetitions produced by speakers are not more than 200 tokens apart from their previous mention.

Locality effects: Same-utterance self-repetitions. We investigate the interaction between cumulativity and recency by focusing on densely clustered repetitions, produced by a speaker within a single utterance (the median distance between repetitions in the same utterance is 8 words; across turns it is 370.5 words). Table 8.4 shows an example of same-utterance repetition. Repeating a

Sp	RI	RI Utt	Dist	Utterance	$H(S)$	$H(c)$	$FE(c;S)$
A	0	0	-	[...] I think that everyone should have the same opportunities and **I don't think you should be** proud or ashamed of what your you know what your situation is whether you what your what your race is whether you're a woman or a man whether you live from this pl whether you're in this place [...]	4.24	1.90	1.21
A	1	0	80	I well I th I don't think it should **I don't think you should be**	3.40	1.73	1.40
A	2	0	19	Well yes perhaps but **I don't think you should be** like um embarrassed about it or I think I think you should just sort of	3.95	1.06	2.25

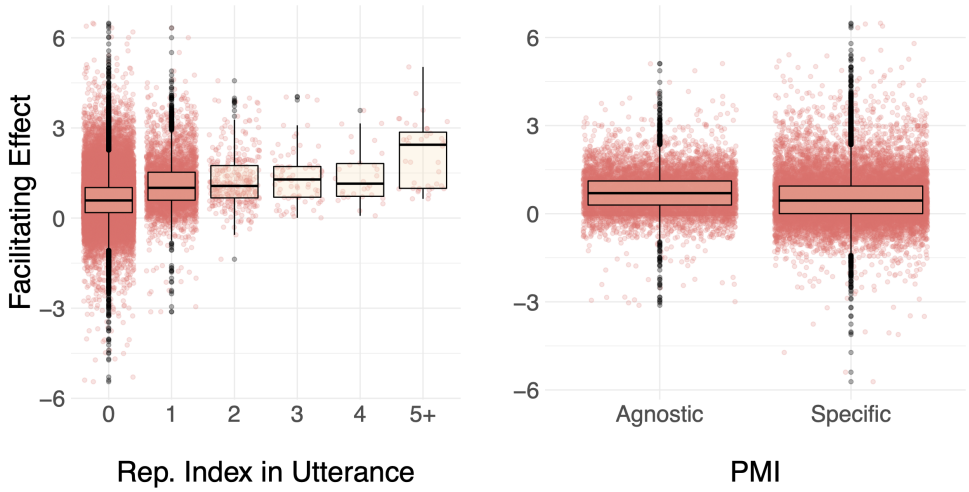
Table 8.5: Repetition chain for the construction *'I don't think you should be'* in dialogue S2AX of the Spoken BNC, annotated with speaker identifier (Sp), repetition index (RI), RI in utterance (RI Utt), and distance from previous mention (Dist; in tokens). $H(u)$ is the utterance surprisal, $H(c)$ and $FE(c;u)$ are the construction's surprisal and facilitating effect.

construction when it has already been mentioned in the current utterance limits its facilitating effect ($\beta = -0.099, p < 0.05$, *95% c.i.* $-0.184 : -0.013$): if a portion of the utterance already consists of a construction, utterance surprisal will already be reduced, which in turn reduces the potential for the facilitating effect of repetitions. Nevertheless, we find **strong cumulativity effects for self-repetitions within the same utterance**: the repetition index *within the current utterance* of a construction mention (i.e., how often the construction has been repeated so far in the utterance) has a positive effect on FE ($\beta = 0.178, p < 0.005$, *95% c.i.* $0.130 : 0.226$); see Figure 8.3a. In sum, same-utterance self-repetitions, especially those involving three or more mentions in a single utterance, can have a strong reduction effect on utterance surprisal. Although this may seem a simple yet very effective strategy for information rate mitigation, it is unlikely to be very effective in terms of the amount of information exchanged. Indeed, speakers do not use this strategy often in the Spoken BNC: only 6.82% of the total construction occurrences have at least one previous mention in the same utterance.

Interaction-specificity. To distinguish interaction-specific constructions—i.e., those repeated particularly often in certain dialogues—from interaction-agnostic ones, we measure the association strength between a construction c and a dialogue d as the pointwise mutual information (PMI) between the two:

$$\mathrm{PMI}(c, d) = \log_2 \frac{P(c|d)}{P(c)} \tag{8.3}$$

This quantifies how unusually frequent a construction is in a given dialogue, compared to the rest of the corpus. For example, for a construction to obtain a PMI score of 1, its probability given the dialogue $P(c|d)$ must be twice as high as its prior probability $P(c)$. Low PMI scores (especially below 1) characterise interaction-agnostic constructions, whereas higher PMI scores indicate that con-



(a) (b)

Figure 8.3: Facilitating effect against repetition index *within the current utterance* (a) and facilitating effect of interaction-agnostic constructions ($\text{PMI}(c,d) < 1$) vs. interaction-specific constructions ($\text{PMI}(c,d) = \max_{c',d'} \text{PMI}(c',d')$) (b).

structions are specific to a given dialogue. The probabilities in Equation 8.3 are obtained via maximum likelihood estimation over the analysis split of the Spoken BNC. PMI scores have a negative effect on *FE* ($\beta = -0.139, p < 0.005$, 95% c.i. $-0.154 : -0.124$), indicating interaction-agnostic constructions have a stronger facilitating effect than interaction-specific ones. Figure 8.3b shows *FE* distributions for the most extreme cases: constructions with a PMI lower than 1 ('agnostic') and constructions that have been repeated in only one dialogue ('specific').

8.6.3 Types of construction

In this section, we analyse factors shaping the facilitating effect of construction forms, rather than individual mentions. We focus on the length of a construction and on whether it is referential.

Construction length has a positive effect on *FE* ($\beta = 0.098, p < 0.005$, *95% c.i.* $0.087 : 0.119$): **longer constructions have stronger facilitating effect.** Tables 8.4 and 8.5 show repetition chains for constructions of length 3 and 6. Non-construction sequences display an opposite, weaker trend ($\beta = -0.019, p < 0.05$, *95% c.i.* $-0.032 : -0.005$), as measured with a linear model. A possible explanation for the positive effect of construction length is related to production cost. Longer constructions are more costly for the speaker, so for them to still be an efficient production choice, their facilitating effect must be higher.

Finally, we observe that **referential constructions have a stronger facilitating effect than non-referential ones**. Our LME model yields a positive effect for referentiality on *FE* ($\beta = 0.124, p < 0.005$, 95% c.i $0.099 : 0.149$) and we find

a significant difference between the *FE* of the two types ($t = 3.115, p < 0.005$, *95% c.i.* $0.072 : 0.016$). Looking in more detail, first mentions of referential constructions have higher surprisal and lower *FE* than first mentions of non-referential ones (*H*: $t = 15.435, p < 0.005$, *95% c.i.* $1.115 : 0.864$; *FE*: $t = -9.315, p < 0.005$, *95% c.i.* $-0.246 : -0.161$), perhaps since words in referential sequences tend to be less frequent and more context-dependent. However, when repeated, their surprisal drops more substantially, reproducing inverse frequency effects attested in humans for syntactic repetitions (Bock, 1986; Scheepers, 2003). As a result, their *FE* exceeds that of non-referential constructions (*FE*: $t = 8.818, p < 0.005$, *95% c.i.* $0.117 : 0.183$), with the surprisal of a repeated reference being 81% lower than that of its utterance context. Overall, these findings indicate that although referential constructions are less frequent than non-referential ones (23.3% vs. 76.7%), their repetition is a particularly effective strategy of information rate mitigation.

8.7 Discussion and conclusions

Construction repetition is a pervasive phenomenon in dialogue; their frequent occurrence gives constructions a processing advantage (Conklin and Schmitt, 2012). In this study, we showed that the processing advantage of constructions can be modelled as reduced surprisal and proposed that speakers' production of constructions can be seen as a strategy for information rate mitigation. This strategy can explain why utterance surprisal is often found to decrease over the course of dialogues, as described in Chapter 7, in contrast with the predictions of theories of optimal use of the communication channel (Genzel and Charniak, 2002).

We observed that, as predicted, construction usage in English open-domain spoken dialogues mitigates the information rate of utterances. Furthermore, while constructions are produced at a stable rate throughout dialogues, their facilitating effect—our proposed measure of reduction in utterance surprisal—increases over time. We found that this increment is led by construction repetition, with facilitating effect being positively affected by repetition frequency, density, and by the contents of a construction. Repetitions of referential constructions reduce utterance surprisal more aggressively, arguably making them a more cost-reducing alternative to the shortening strategy observed in chains of referring expressions (Krauss and Weinheimer, 1964, 1967), which instead tends to preserve rate constancy (as seen in Chapter 7).[11]

Relation to cognitive effort. We consider repetitions as a way for speakers to make dialogic interaction less cognitively demanding both on the production and on the comprehension side. This is not at odds with the idea that repetitions are driven by interpersonal synergies (Fusaroli et al., 2014) and coordination (Sinclair

[11] Expression shortening ca be more efficient, however, in terms of articulatory cost.

and Fernández, 2021). The operationalisation of these higher level processes can be described by means of lower level, efficiency-oriented mechanisms, with synergy and coordination both corresponding to reduced collaborative effort. Although surprisal estimates from neural language models have been shown to be good predictors of processing effort, measured as reading time, gaze duration, and N400 response (van Schijndel and Linzen, 2018; Wilcox et al., 2020; Meister et al., 2021; Merkx and Frank, 2021), we cannot claim that our work directly models human cognitive processes as we lack the relevant human data to measure such correlation for the corpus at hand.

Adaptive language model. Our decision to use an adaptive neural language model affects surprisal estimates in two main ways. On the one hand, due to their high frequency, constructions are likely to be assigned higher probabilities by this model, and therefore lower surprisal. For this reason, we did not present constructions' lower surprisal as a novel result. As explained in Section 8.5.2, this is a precondition for our experiments on the facilitating effect of constructions, which is not determined exclusively by their surprisal (as empirically shown in Section 8.5.4) but rather measures the effect of construction usage on the surprisal of entire utterances. On the other hand, because our model is adaptive, the probability of constructions is likely to increase as a result of their appearance in the dialogue history. Adaptation, however, also contributes to lower utterance surprisal *overall* through the exploitation of topical and stylistic cues, as demonstrated by the lower perplexity of the adaptive model on the entire target dialogue as well as on other dialogues from the same dataset (see Section 8.4.2 and Appendix B.6.2). In sum, while our adaptive language model assigns higher probabilities to frequently repeated tokens—as expected from a psychologically plausible model of utterance processing—it is not responsible for the discovered patterns of construction facilitating effect. In future work, the model can be improved, e.g., by conditioning on the linguistic experience of individual speakers.

Types of dialogue. To consolidate our findings, construction repetition patterns should also be studied in dialogues of different genres and in datasets where utterance surprisal was not found to decrease. We haven chose the Spoken BNC as it contains dialogues from a large variety of real-life contexts, which makes it a representative dataset of open-domain dialogue. In task-oriented dialogue, we expect constructions to consist of a more limited, task-specific vocabulary, resulting in longer chains of repetition and potentially more frequent referential construction usage. These peculiarities of task-oriented dialogue may influence the strength of the facilitating effect (as we have seen, facilitating effect is affected by both frequency and referentiality) but we expect our main results to still hold, as they are generally related to the processing advantage of constructions.

Relevance for dialogue generation models. Besides contributing new empirical evidence on construction usage in dialogue, our findings inform the development of more naturalistic utterance generation models. They suggest that models should be continually updated for their probabilities to better reflect human expectations; that attention mechanisms targeting contexts of different sizes (local vs. global) may have a significant impact on the naturalness of generated utterances; and that while anomalous repetitions (e.g., generation loops) should be prevented (Li et al., 2016; Holtzman et al., 2019), it is important to ensure that natural sounding repetitions are not suppressed. We expect dialogue systems that are able to produce human-like patterns of repetitions to be perceived as more natural overall—with users having the feeling that common ground is successfully maintained (Pickering and Garrod, 2004)—and to lead to more effective communication (Reitter and Moore, 2014). In our view, such human-like patterns can be reproduced by steering generation models towards the trends of information rate observed in humans.

Part Three

Utterance Production

The studies presented so far in this thesis build on the idea that having a computational model of language comprehension allows to study how a hypothetical comprehender's interpretation and processing affect speakers' usage of words and utterances. To mimic the comprehension process, neural language models are employed in these studies as 'recognisers', i.e., to compute representations of and assign probabilities to natural language strings. We now move from neural models as recognisers to neural language 'generators', i.e., systems from which natural language strings can be sampled. This will allow us to obtain artificial simulations of speakers' utterance production process, and in turn, as we will see especially in Chapter 11, to study aspects of human comprehension behaviour. Hopefully, the rationale behind the titles chosen for the second and third part of the thesis, which can be summarised as *'production through comprehension'* and *'comprehension through production'*, is now more perspicuous.

While using neural language models to study language comprehension is a rather established approach, whose successes I have discussed at length in Part 2, the approach I take in this part of the thesis—using neural text generators as models of language production—is less conventional and requires validation. Part 3 will therefore begin with an evaluation of neural text generators as statistical processes of utterance production. We will assess generators' compliance to the statistical properties of the human language process, empirically estimated via samples from human populations. After obtaining this validation, neural text generators will be used as models of utterance production, in order to test aspects of comprehension behaviour known to be affected by comprehenders' expectations over speaker utterances. Part 3 will then end with a position piece on how to design computational models of language production which more accurately and reliable mimic human language use.

Producing language is making decisions about which bits of behaviour—or signals—to use in order to reduce the uncertainty of a comprehender over the space of possible future states of the environment (Rosenberg and Cohen, 1964; Levy, 2008b; Goodman and Frank, 2016). Production is successful when the chosen linguistic behaviour restricts the space of possibilities to the environment states that correspond to the speaker's communicative intent. Uttering a word, for example, may cut the space of possibilities in half, while uttering a whole sentence may restrict the comprehender's uncertainty to, say, $1/64$ of the state space—thus proving more informative. This intuitive representation of the amount of information conveyed by a signal has an elegant probabilistic characterisation, which has been at the core of the studies in Part 2. The probability of a signal that cuts the space of possibilities into I parts is $p = (1/2)^I$, which is equivalent to saying

that the amount of information carried by that signal is $I = -\log_2(p)$.[12] This is the signal's information content or surprisal. As we have seen previously, its unit of measure is the *bit*: a signal that restricts uncertainty to $1/64$ of the state space has a probability of $(1/2)^6 = 1/64$ and its surprisal is $-\log_2(1/64) = 6$ bits.

When the signal is a complex form of behaviour such as a natural language utterance, and the state space is the space of natural language strings, this probabilistic characterisation of uncertainty becomes problematic. The state space is high-dimensional, structured, and unbounded, so utterance surprisal must be computed by aggregating token-level estimates, typically obtained from autoregressive language models (e.g., Meister et al., 2021; Wallbridge et al., 2022, and our work in Part 2). As a result, different realisations of the same communicative intent (for example, *'I have to go'* and *'I gotta go'*) compete for probability mass (Holtzman et al., 2021), and different dimensions of uncertainty are conflated into a single quantity, the signal's probability (and thus surprisal). This makes it difficult to appreciate whether the information carried by an utterance results from the unexpectedness, for example, of its lexical material, syntactic arrangements, semantic content, or speech act type (Arehalli et al., 2022).

To tackle this open problem, in Chapter 10, we design probes that provide a low-dimensional view of uncertainty over sequence-level language model outputs. Our probes allow for an interpretable and actionable quantification of uncertainty: they measure uncertainty at an instance-level, in terms of the lexical, syntactic, and semantic similarity between utterances produced (by humans or models) given an individual linguistic context. This chapter also defines a statistical framework to assess the alignment between the language model's representation of sequence-level uncertainty against plausible response variability in humans and includes an empirical study on four natural language generation tasks.

In Chapter 11, we build on this approach to uncertainty quantification and take inspiration from the concept of *alternatives* in semantics and pragmatics (Horn, 1972; Stalnaker, 1978; Gazdar, 1979; Grice, 1975; Levinson et al., 2000; Falaus, 2013, i.a.) to obtain estimates of what we will call the *'information value'* of an utterance. Information value is a new intuitive representation of the predictability of utterances which is unaffected by the issues described above for surprisal. It is simply the distance of an utterance from the set of alternative productions that are expected by a comprehender. Our study will show that empirical estimates of information value obtained using neural text generators can predict and explain human reading times and acceptability judgements in dialogue and text.

Part 3 is concluded by Chapter 12, which is a reflection on what current neural text generators lack, and how they can be improved, when it comes to reproducing human-like strategies of language use. In particular, this chapter collects insights from the studies presented so far in the thesis into a formal

[12] $(\frac{1}{2})^I = p \Leftrightarrow 2^I = \frac{1}{p} \Leftrightarrow I = \log_2(\frac{1}{p}) \Leftrightarrow I = -\log_2(p)$.

framework to characterise efficient and communicatively effective strategies of utterance production, and to implement them in Natural Language Generation systems. The main argument that it will bring across is that human-like linguistic behaviour emerges as a result of reasoning about context, goals, costs, and utility, and it discusses promising directions for natural language generation research that takes these concepts seriously.

Chapter 9

Background

9.1 Natural language generation

The goal of natural language generation (NLG) research is to give computers the ability to mimic human linguistic communication. NLG systems can be trained to perform a wide variety of language tasks, ranging from translation to storytelling, and they can interact with humans in a conversational way to help them solve real-world problems. As NLG models become more accurate and versatile, it is of paramount importance to design evaluation procedures that guarantee the models are safe and trustworthy.

Evaluation is often performed with automatic metrics, including measures of overlap between candidate generations and human references, and via models trained directly on human judgements to predict the perceived quality of new generation candidates. Chapter 10 will focus on another important dimension: the models' uncertainty over possible generations. The proposed approach, complementary to other types of automatic evaluation, makes model assessments particularly reliable because it does not judge a model only by a single output, but also by what it could have generated.

This section contains a discussion of automatic approaches to the evaluation of NLG systems, which will serve as a background for Chapter 10.

9.1.1 Automatic approaches to NLG evaluation

Evaluation of NLG systems is an important research area in NLP with at least two main strands. On the one hand, **automatic evaluation** approaches are of high practical importance for model selection and quality-aware decoding algorithms (Borgeaud and Emerson, 2020; Eikema and Aziz, 2020; Fernandes et al., 2022; Suzgun et al., 2022). On the other hand, **human evaluation** plays a crucial role in assessing systems (Belz and Gatt, 2008) as well as automatic evaluation metrics (Reiter, 2018). Efforts in this area are the result of a long-

standing yet increasing awareness of the limitations of existing evaluation approaches (Callison-Burch et al., 2006; Dušek et al., 2020), of drastic improvements in NLG systems—whose output can often no longer be distinguished from human productions (Gehrmann et al., 2022; Dou et al., 2022)—and of the increase in the number, variety, and open-endedness of the tasks that modern NLG systems are intended to model (e.g., See et al., 2019; FAIR Diplomacy Team et al., 2022; Schick et al., 2022).

Reference-based. The most common way of automatically evaluating text generators is via metrics that estimate the overlap between candidate generations and references (e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), BertScore (Zhang et al., 2020a), i.a.). Reference-based metrics are less suited for open-ended tasks such as story generation and dialogue (Liu et al., 2016), where a single reference (or even a handful) cannot be representative of the large space of plausible communicative goals and realisations.

Reference-free. A popular, reference-free alternative is to train evaluation models that discriminate human productions from model outputs (e.g., Bruni and Fernández, 2017; Gehrmann et al., 2019; Hashimoto et al., 2019), score the appropriateness of input-output pairs (e.g., Sinha et al., 2020; Fomicheva et al., 2020), or model human judgements directly (Lowe et al., 2017; De Mattei et al., 2021; Rei et al., 2021). Neural language models themselves have been recently proposed as evaluators (Yuan et al., 2021), and have sometimes been used to assess generations along interpretable evaluation dimensions (Deng et al., 2021; Zhong et al., 2022). Reference-free metrics of this kind have been criticised for being inherently biased toward models which are more similar to the evaluator and thus limited in their ability to evaluate generated text (Deutsch et al., 2022).

Statistical evaluation. Statistical evaluation compares model generations to human productions *in distribution* through real-valued statistics (e.g., Zipf's coefficient, type token ratio, length) as opposed to strings themselves. These statistics are typically compared marginally at the corpus level (Eikema and Aziz, 2020; Meister and Cotterell, 2021; Pillutla et al., 2021; Pimentel et al., 2022), supporting general claims about model performance in relation to humans. More recently, Barkhof and Aziz (2022) and Deng et al. (2022) compared statistics at the instance level, supporting claims about models' performance in relation to humans for individual inputs. In Chapter 10, we craft statistics that evaluate models' uncertainty at the *instance level* against the variability *over sequences* observed in multi-reference NLG datasets. Although evaluating uncertainty is gaining traction in NLP (e.g., Desai and Durrett, 2020; Glushkova et al., 2021; Baan et al., 2022), there is relatively little work on sequence-level uncertainty

estimation and evaluation (Ott et al., 2018; Malinin and Gales, 2020; Aina and Linzen, 2021; Kuhn et al., 2023).

9.2 Expectations, predictability, and surprisal

In Chapter 11, neural text generators will be used to model expectations over plausible upcoming linguistic units. To better understand the contributions of that chapter, a second important background topic to discuss is the relation between expectations, predictability, and surprisal within linguistic theories of language processing.

Expectation-based theories of language processing define the effort required to process a linguistic unit as a function of its predictability. Surprisal theory, perhaps the most prominent example, posits a direct relationship between effort and predictability, quantified—well—as surprisal (Hale, 2001). The theory is supported by broad empirical evidence across domains and languages (de Varda and Marelli, 2022; Pimentel et al., 2021; Wilcox et al., 2023), and serves as a foundation for quantitative principles of language production and comprehension such as the Entropy Rate Constancy (ERC; Genzel and Charniak, 2002) and the Uniform Information Density (UID; Levy and Jaeger, 2007) principles studied in Part 2.

Without direct access to the conditional probabilities of linguistic units, psycholinguists have relied on statistical models of language such as Markov chains to estimate surprisal (Hale, 2001; McDonald and Shillcock, 2003). More recently, large-scale language models have emerged as powerful estimators of token-level surprisal, thanks to their ability to predict different aspects of human language comprehension behaviour (what is often referred to as their *psychometric predictive power*). Psychometric variables include self-paced and eye-tracked reading times (Keller, 2004; Goodkind and Bicknell, 2018b; Wilcox et al., 2020; Meister et al., 2021; Shain et al., 2022; Oh and Schuler, 2022), acceptability judgements (Lawrence et al., 2000; Heilman et al., 2014; Lau et al., 2015, 2017; Warstadt et al., 2019; Wallbridge et al., 2022), and brain response data (Frank et al., 2015; Schrimpf et al., 2021).

To obtain estimates of utterance surprisal, different aggregates of token-level surprisal have been proposed, motivated by psycholinguistic theories like ERC and UID. However, their behaviour is far less understood. For example, divergences between how model characteristics affect predictive power for different comprehension tasks (Meister et al., 2021) and across languages (Wilcox et al., 2023) raise questions about whether token-level aggregates appropriately capture human utterance processing. In Chapter 11, I will present an sequence-level alternative to utterance surprisal with complementary, and in some cases superior, psychometric predictive power.

Chapter 10

Evaluating uncertainty in neural text generators

The content of this chapter is based on the following preprint, which is currently under submission:

> Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What Comes Next? Evaluating Uncertainty in Neural Text Generators Against Human Production Variability. To appear in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Republic of Singapore. Association for Computational Linguistics.

Mario and Joris jointly produced the research idea. Mario and Joris performed the experiments and wrote most of the article; Wilker also contributed to the writing. Wilker, Raquel, and Barbara provided advice throughout the project, then reviewed and revised the manuscript. The text in this chapter overlaps with that of the online preprint.

10.1 Introduction

Humans display great variability in language production, in particular when the context or the task are open-ended—such as in storytelling or in dialogue. Given a story prompt, for example, there are many plausible ways in which different humans (or a single writer, if asked multiple times) may tell the story (Fan et al., 2018). We refer to this phenomenon as *production variability*. Production variability in humans has two main sources. First, when situated in a context, speakers may have variable communicative goals (Austin, 1975; Searle, 1969; Sacks et al., 1974), and the number and variety of plausible communicative goals depends on the production task (Jokinen, 1996). Translation, for instance, defines the communicative goal almost unequivocally while a dialogue context might allow for a wide variety of communicative goals (expressed, e.g., as a request, an assertion, or a yes-no question). The second source of variability is the fact that even when context and communicative goal are fixed, speakers' linguistic realisations of the communicative goal may vary (Levelt, 1993). Both sources of variability apply to individuals as well as to populations: if an expert is asked to simplify a complicated sentence multiple times, they may perform different rewriting transformations (e.g., paraphrasing, reordering, or sentence splitting) and produce different texts (Alva-Manchego et al., 2021); the same is true if multiple experts are asked to perform the task (Xu et al., 2015).

If we are to regard a natural language generation (NLG) system (or *text generator*) as a good model of human language production, it should capture the variability observed in humans. A text generator combines two mechanisms: (i) an underlying statistical model—typically, an autoregressive factorisation of the probability of sequences, with conditional token probabilities predicted by a neural network; and (ii) an iterative decoding algorithm that chains samples from next token distributions into a complete production. Together these two mechanisms specify a probability distribution over sequences of tokens, which can be regarded as a representation of the model's uncertainty (Halpern, 2017) about productions for a given generation context. In this work, we assess whether this representation of uncertainty is in compliance with production variability exhibited by a population of humans—which in turn, we argue, can be regarded as an expression of *aleatoric* uncertainty, i.e., irreducible uncertainty due to the stochastic nature of the data generating process (Der Kiureghian and Ditlevsen, 2009; Hüllermeier and Waegeman, 2021).

Quantifying the closeness in distribution between a text generator and a human population is difficult for multiple reasons: we only have an iterative view into the generator's distribution; the 'human distribution' is an implicit or even hypothetical object; and in both cases, the sample space is large or even unbounded. We can, however, compare these two objects via aspects of the samples they produce and assess their statistical similarity—which is what we propose in this study. For each individual generation context, we compare scalar properties

of generations (through repeated model sampling) and human productions (using multi-reference NLG datasets). In particular, we *probe* for lexical, syntactic, and semantic similarity between productions, thus allowing for a quantitative and interpretable assessment of uncertainty.

We find that the uncertainty of neural text generators is higher than justified by human production variability in open-ended tasks, like story generation and open-domain dialogue; and that it is lower on more constrained tasks, like machine translation and text simplification. Popular decoding algorithms, that bias away from the distribution of the generator's underlying statistical model (e.g., top-k or top-p, rather than ancestral sampling), have a limited impact on the generator's ability to faithfully represent human variability. We complement our quantitative assessments with a detailed analysis of individual generation contexts, which sheds light on whether a generator has robustly learned to reproduce degrees and aspects of human variability plausible for the communicative task.

Our work has important implications for NLG evaluation and data collection. Multiple samples and, when possible, multiple references, should be used to assess the statistical fitness of text generators. This approach, complementary to other types of automatic evaluation, makes model assessments particularly insightful and trustworthy because it does not judge a model only by a single output, but also—intuitively—by *what it could have generated*, and it provides tools for error analysis at the instance-level along a dimension of the data generation process that is often neglected.

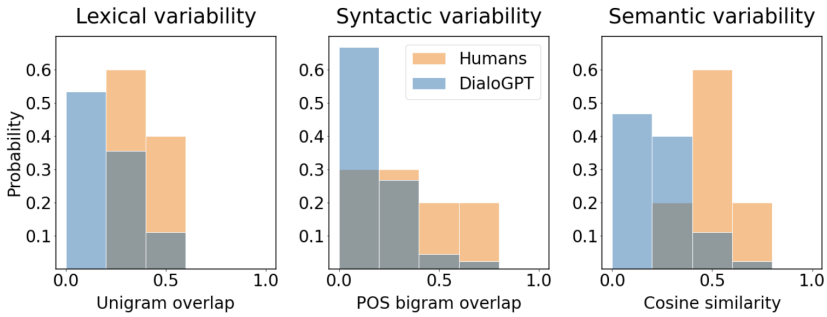
10.2 Probing language processes for production variability

We interpret language production—by humans or text generators—as captured by a probability distribution over productions, a random variable Y, given a linguistic context $X = x$. The context x can be a source sentence in machine translation, a story prompt in story generation, or more generally the *input* to a language process. In turn, a production is a piece of text y such as a single translation, a story, or more generally the *output* of a language process.[1]

10.2.1 Production variability

For any language process, *production variability* is fully characterised by a conditional probability distribution $p_{Y|X=x}$ representing uncertainty about the output Y given input $X = x$. However, analysing this distribution is difficult. Notably,

[1]**Notation.** Random variables are denoted by uppercase letters (e.g., Y) and outcomes by lowercase letters (e.g., y); $p_{Y|X=x}$ denotes the probability distribution of Y given $X = x$ and $p_{Y|X=x}(y)$ its probability mass function.



Dialogue context

It's very dark in here. Will you turn on the light?
Okay. But our baby has fallen asleep.
Then, turn on the lamp, please.
But where's the switch?

Humans

- Lamp is upon the study table and now you know where the switch is.
- I will light up the torch, so you can find the switch and turn on the lamp.
- Don't you know where the switch is?
- Switch is on the left side of the lamp.
- Just press the second switch on the board.

DialoGPT-medium, nucleus $p = 0.9$

- You don't have one.
- Where's the button?
- I have a few, try and figure it out.
- It's on the top.
- Well, you'll want to turn it on.
- Turn it on.
- I'm sorry.
- On my chest
- I'm on it!
- Turning on the switch

Figure 10.1: Production variability observed in 5 human responses vs. 10 responses generated by DialoGPT. Each graph presents the distribution over a similarity metric. The generated responses exhibit higher lexical, syntactic, and semantic variability than humans'. The generator's uncertainty is too high in this dialogue context.

for human language processes, we do not have an explicit representation of this object, but through data collection we can draw conditional samples from it (i.e., gather references given a context). For NLG models, we do have an algorithmic representation, which is usually sufficient to enable sampling.

Analysing language processes through their samples introduces other difficulties as text is a high-dimensional, structured, non-numerical data type. For tractable analysis, we exploit a set of real-valued and interpretable statistics, or *production probes*, to re-express a language process' distribution in terms of how, given an input, its outputs relate to outputs of another language process. When both processes are independent humans performing a task, we obtain a sense of how plausible human productions relate (or *vary* with respect) to other plausible human productions, along a linguistically interpretable dimension. When

we swap one or both processes for an NLG system, we obtain tools to analyse how model generations relate to plausible human productions, thus assessing a model's representation of uncertainty against the variability observed in humans.

Specifically, given a context x, two language processes with distributions $p_{\hat{Y}|X=x}$ and $p_{Y|X=x}$ and a choice of similarity metric $k(\cdot,\cdot) \in \mathbb{R}$, our probe for production variability is a real random variable $k(\hat{Y}, Y)$ which captures the joint distribution of pairwise similarity between any two outputs drawn conditionally from the language processes. The exact distribution of $k(\hat{Y}, Y)$ is intractable to characterise, but we can estimate it via simulation by drawing from the two processes and assessing the similarity metric on sampled pairs (Figure 10.1). Consider the case where we analyse the human language process (Section 10.4) through $k(Y, Y)$: when multiple realisations of the output are dissimilar—e.g., 'Fantastic, thank you!' and 'I asked you first' given the input 'How is your day?'—production variability is high along the dimension captured by k.

10.2.2 Production probes

To assess the closeness of random draws from two language processes given the same input context, we instantiate our production probes with three similarity functions: lexical, syntactic, and semantic. There are many other dimensions one may be interested in and we encourage future work to find the production probes that best capture them.

Lexical. The fraction of common n-grams in two texts, with $n \in [1, 2, 3]$ (i.e., number of matching n-gram occurrences divided by the total number of n-grams in both strings).

Syntactic. The fraction $\text{syn}_n(y', y)$ of common part-of-speech (POS) n-grams in two texts (i.e., number of matching POS n-gram occurrences divided by the total number of POS n-grams in both strings).[2]

Semantic. Cosine similarity $\text{sem}(y', y)$ between the sentence embeddings of two texts (Reimers and Gurevych, 2019).[3]

10.3 Experimental setup

We experiment with four NLG datasets that contain 5 or more human references per input instance and for which we expect humans to display different degrees of production variability. Table 10.1 shows relevant statistics. All datasets are in English; for translation, the target language is German. The reference collection

[2]We use the `en_core_web_md` model from spaCy (Honnibal et al., 2020).

[3]`sentence-transformers/all-distilroberta-v1`

procedure varies across datasets. We discuss how this may impact our analysis in Section 10.7.

Machine translation. We use 500 sentences from the WMT-14 En-De test set (*newstest2014*; Bojar et al., 2014), which have been annotated by Ott et al. (2018) with 10 additional reference translations produced by as many human translators, and As a generator, we use Helsinki-NLP's Transformer-Align model trained on Opus-MT (Tiedemann and Thottingal, 2020).

Text simplification. We use the 2,000 instances of the *ASSET* validation set (Alva-Manchego et al., 2020). For each source sentence, originally from the TurkCorpus (Xu et al., 2016), ASSET includes 10 additional simplifications by as many crowdsource annotators. On this dataset, we test Flan-T5 Large (Chung et al., 2022), an instruction fine-tuned version of the T5 language model (Raffel et al., 2020), further fine-tuned on the ASSET training set.

Storytelling (Story generation). We use the 759 instances from the *WritingPrompts* test set (Fan et al., 2018) for which at least 5 human references are available. Prompts and stories are originally scraped from `r/WritingPrompts`, a Reddit forum of stories written by online users in response to story prompts designed by other users.[4] The number of stories available per prompt (9.56 ± 7.67) varies from 5 to 92. As a generator, we use GPT-2 Large (Radford et al., 2018) fine-tuned on the WritingPrompts training set.

Open-domain dialogue. We use the development set of *DailyDialog++* (Sai et al., 2020), which contains 5 additional references for 1028 conversations from the DailyDialog corpus (Li et al., 2017). The dialogues are short (less than 8 turns) and cover a broad list of topics; for each dialogue, 2-3 annotators were asked to generate 1-3 alternative responses.[5] For this task, we use the pretrained and DialoGPT Medium (Zhang et al., 2020c).

10.3.1 Decoding algorithms

Candidate generations can be sampled directly from a text generator one word at a time, via (i) *unbiased sampling* (also 'ancestral' or 'forward' sampling; Bishop, 2006; Koller and Friedman, 2009). Most often, though, other slightly more complex decoding algorithms are used to obtain candidates from a generator. We

[4] `https://www.reddit.com/r/WritingPrompts/`

[5] The *DailyDialog++* annotators are also instructed to avoid short generic responses such as '*Sure*' and to write, instead, meaningful responses with at least 8-10 words.

		Machine Translation		*Text Simplification*		*Story Generation*		*Open-Domain Dialogue*	
		MEAN ± STD	RANGE	MEAN ± STD	RANGE	MEAN ± STD	RANGE	MEAN ± STD	RANGE
Input	Words	23.34 ± 11.35	3-67	22.26 ± 8.92	7-57	25.40 ± 14.18	1-68	47.62 ± 30.37	5-311
	Tokens	25.79 ± 12.91	4-81	28.00 ± 11.68	7-78	26.49 ± 14.68	1-70	48.94 ± 31.52	5-326
	Sentences	1.01 ± 0.09	1-2	1.02 ± 0.14	1-2	1.75 ± 0.93	1-6	5.49 ± 2.82	1-22
	Words in sent.	23.15 ± 11.37	2-67	21.80 ± 9.11	1-57	14.48 ± 7.89	1-50	8.67 ± 5.20	1-50
	Tokens in sent.	25.58 ± 12.90	2-81	27.42 ± 11.88	1-78	15.12 ± 8.13	1-51	8.93 ± 5.39	1-50
Output	Words	21.96 ± 10.99	2-66	19.57 ± 8.29	4-62	659.72 ± 450.46	101-2,681	10.61 ± 4.85	2-46
	Tokens	27.28 ± 14.09	5-86	24.22 ± 10.65	5-91	696.66 ± 476.93	104-2,961	10.84 ± 5.01	2-53
	Sentences	1.06 ± 0.25	1-4	1.33 ± 0.56	1-5	47.76 ± 35.44	1-308	1.32 ± 0.52	1-5
	Words in sent.	20.67 ± 10.86	1-66	14.70 ± 6.71	1-59	13.81 ± 9.59	1-722	8.06 ± 4.32	1-36
	Tokens in sent.	25.69 ± 13.92	1-86	18.19 ± 8.78	1-91	14.63 ± 10.22	1-722	8.24 ± 4.45	2-37

Table 10.1: Length statistics in number of tokens, as obtained with the tokenisers of the language models used for generation (Section 10.3).

experiment with (ii) *temperature sampling*, i.e. unbiased sampling after a renormalisation of the output space (lower temperatures α make the model increasingly confident); (iii) *top-k sampling* (Fan et al., 2018), which limits the sampling distribution to the k most likely words at each time step; (iv) *nucleus sampling* (Holtzman et al., 2019), which dynamically limits the sampling distribution to the smallest vocabulary subset whose cumulative probability exceeds a threshold p; (v) *locally typical sampling* (Meister et al., 2023), which sorts words based on the deviation of their negative log probability from the expected conditional entropy of the distribution, and redefines the sample space by taking words from the top of this sorted list until their cumulative probability exceeds a threshold τ. For all decoding algorithms, we set the maximum generated sequence length to 100 (cf. Table 10.1).

10.4 Human production variability across NLG tasks

Consider $p_{Y|X=x}$ the distribution that describes the human language process, and define the following special case:

$$H_k(x) \coloneqq k(Y, Y) \ . \tag{10.1}$$

Estimating this probe by drawing pairs of human productions provides an interpretable view on plausible variability along the dimension captured by k. Figure 10.2 shows $H_k(x)$ marginalised over inputs for the four NLG tasks. High similarity indicates low variability, and vice versa. We use unigram overlap for the lexical probe, POS bigram overlap for the syntactic probe, and cosine similarity for the semantic probe.

Translation and text simplification. Humans show low production variability in these two tasks. While translations of a given source sentence are more lexically and semantically varied, simplifications exhibit a higher degree of syntactic variability, probably as a result of the instructions used during data collection (writers were asked to use varying rewriting transformations). Overall, low levels of variability are to be expected as, in both tasks, content preservation is part of the communicative goal.

Story generation. Variability in story generation is strongly dependent on the production probe type. It is low at the syntactic level—close to translation and simplification—while lexical and semantic similarity probes place this task closer to open-domain dialogue. Stories generated from a given prompt may vary a lot in content, but basic syntactic structures and some lexical material are shared. While this task can be a priori perceived at least as 'open-ended' as dialogue, the lower levels of variability may be a result of contextual factors specific to the *WritingPrompts* dataset that we are not explicitly modelling, such as writers being able to read stories contributed by other users.

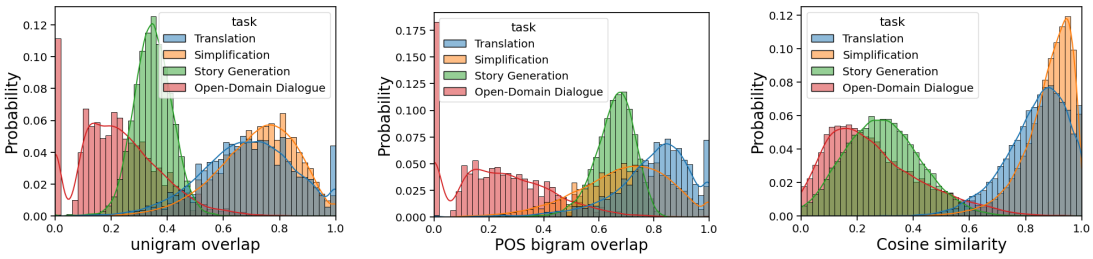
Open-domain dialogue. We observe the highest production variability in this task across all similarity probes. Many output pairs are lexically and syntactically completely dissimilar, as indicated by the zero-bin in Figures 10.2a and 10.2b. Lexical variability is even more extreme when looking at bigrams and trigrams (Figures C.1-C.2 in Appendix C.1) suggesting that while responses rarely share words or phrases, they still sometimes convey similar meaning (Figure 10.2c). Overall, the fact that dialogue appears to be the most open-ended task can be explained by the wide variety of communicative goals that can plausibly follow from a dialogue context and, in part, by the fact that individual annotators produced multiple responses for the *DailyDialog++* dataset and thus were able to monitor the diversity of their outputs.

10.5 Neural text generators' compliance to human production variability

Consider, now, a second language process: a text generator with distribution $p_{\hat{Y}|X=x}$. We study this generator's uncertainty about outputs given an input x under two lenses. In Section 10.5.1, we study how outputs vary with respect to one another, similar to how we defined human production variability $H_k(x)$. We refer to this as the generator's *self-variability*:

$$M_k(x) \coloneqq k(\hat{Y}, \hat{Y}) \ . \tag{10.2}$$

In Section 10.5.2, instead, we study how model generations vary with respect to a language process known to be plausible: a human language process $p_{Y|X=x}$. We



(a) Lexical variability (b) Syntactic variability (c) Semantic variability

Figure 10.2: Human production variability across four NLG tasks. The values on the horizontal axis are single samples of lexical, syntactic, or semantic similarity between two productions for each input (see Section 10.2). Probability mass on the right side signals high similarity and low variability, and vice versa. A large spread indicates that production variability varies widely across inputs.

refer to this as *cross-variability*:

$$C_k(x) \coloneqq k(\hat{Y}, Y) \ . \tag{10.3}$$

Our expectation is that generators with a good representation of aleatoric uncertainty reproduce human production variability along these two axes. As we employ a similarity metric, it may look like we should regard a model as a good approximation to the human process whenever $C_k(x)$ concentrates about large positive values. To some extent, this is the interpretation exploited by most automatic evaluation metrics (single- or multi-reference). In this work, we refrain from taking any one human production as a 'reference' to be closely 'matched'; rather, we take statistical properties of human productions as illustrative of plausible variability and thus targets to be reproduced. We quantify deviation from plausible human variability by estimating a notion of statistical divergence $\mathrm{D}(\cdot, H_k(x))$. Concretely, we employ D_{W_1}, the Wasserstein 1-distance, and D_{μ}, the difference between mean pairwise similarities $\mu_{H_k(x)} - \mu_{M_k(x)}$.[6]

10.5.1 The underlying statistical model

In this section, we evaluate the underlying statistical model (as a result of parameter estimation via MLE) using unbiased sampling. As models observe variability only marginally (multiple references are rarely used during training), it is interesting to study the compliance of their self-variability to human variability:

[6] $W_1(\cdot, \cdot)$ quantifies a notion of distance between two probability measures and is particularly convenient for it can be estimated using Dirac deltas (samples from those measures) (Peyré et al., 2019) more easily than alternatives such as Kolmogorov–Smirnov and total variation distance (which require binning the measurements into empirical cdfs/pdfs). $W_1(M_k(x), H_k(x))$ has an interpretation in terms of 'mass' (in units of k) that has to be moved, on average, to transform one set of samples into another.

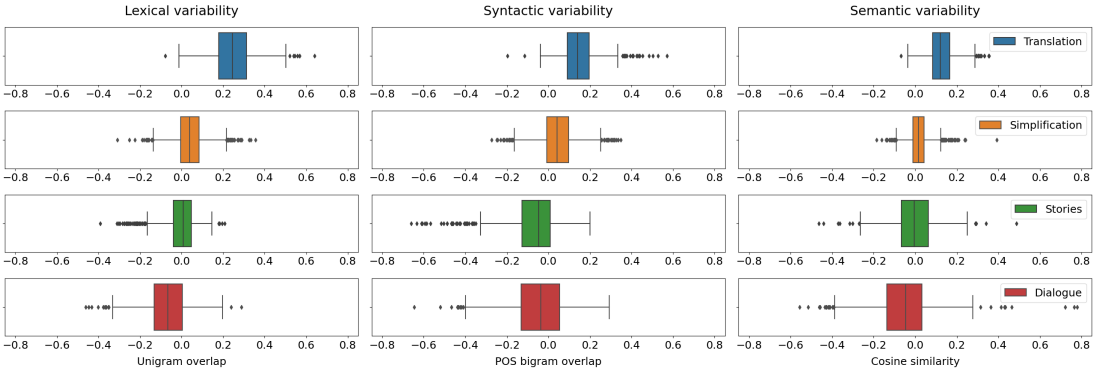


Figure 10.3: Distribution of $\mu_{M_k(x)} - \mu_{H_k(x)}$ over instances. Values greater than zero indicate that the model underestimates the variability of the task (higher mean pairwise similarity); values below zero indicate variability overestimation.

given individual input instances, do similarities between unbiased model samples distribute similarly to similarities between human productions? To distinguish over-estimation from under-estimation, we report a signed notion of divergence, $\mu_{M_k(x)} - \mu_{H_k(x)}$. When $M_k(x)$ and $H_k(x)$ distribute similarly, their mean difference is low for a given x. Negative differences imply that models overestimate variability, i.e., model samples vary more with respect to one another than human samples do; positive differences indicate that models underestimate variability.

Figure 10.3 shows how mean differences distribute across each task-specific test set for the models in Section 10.3 (for other decoding algorithms see Section 10.5.2). We use up to 10 human productions (5 for dialogue) and 10 generations. The first two rows show that $\mathrm{D}_\mu(M_k(x), H_k(x))$ distributes far below 0 for translation (OpusMT) and slightly below 0 for simplification (Flan-T5), indicating that the two models substantially underestimate variability. The opposite is true for dialogue and story generation: both GPT-2 and DialoGPT slightly overestimate the open-endedness of these tasks. Overall, models on more constrained tasks tend to more strongly underestimate variability as captured by $M_k(x)$ while models on open-ended tasks overestimate it slightly.

We also inspect $\mathrm{D}_\mu(C_k(x), H_k(x))$, finding better overall calibration of cross-variability, especially for translation and simplification (see Figure 10.4). That is, when generations are evaluated against productions from a language process known to be plausible, the human language process, models' representation of uncertainty is well aligned to human production variability.

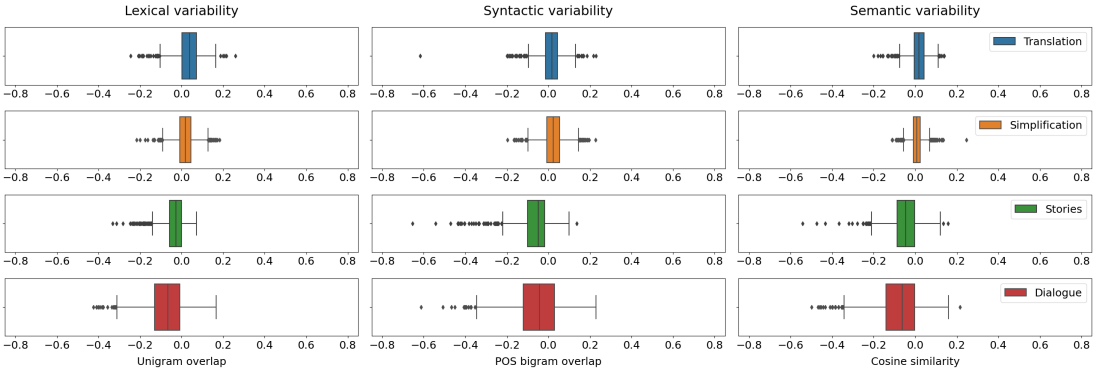


Figure 10.4: Distribution of $\mu_{C_k(x)} - \mu_{H_k(x)}$ over instances. Values greater than zero indicate that the model underestimates the variability of the task (higher mean pairwise similarity); values below zero indicate variability overestimation.

10.5.2 The effect of decoding algorithms

We now study text generators obtained by varying the sampling procedure.[7] We analyse their representation of uncertainty by assessing the divergence between the distribution of generator-human cross-variability $C(x)$ and human variability $H(x)$. While $\mu_{C_k(x)} - \mu_{H_k(x)}$ can inform us about the direction of miscalibration, we observe only a handful of cases where different decoding strategies yield both under- and over-estimation for the same model (Figures C.4 and C.5 in Appendix C.1). Here, we thus report a measure of divergence that is more robust to distributions with multiple modes: the Wasserstein 1-Distance $\mathrm{D}_{W_1}(\cdot, H_k(x))$. Results for self-variability $M(x)$ and mean distance can be found in Appendix C.1.

Human control group. The blue curve in Figure 10.5 shows how the divergence $\mathrm{D}_{W_1}(C_k(x), H_k(x))$ distributes over inputs for unbiased samples from GPT-2 on story generation. To contextualise this observation we report a human control group (the orange curve): this is D_{W_1} measured between two human populations (i.e., we make two disjoint samples from the available human productions for each prompt, use those to estimate $H_k(x)$ and an analogous $\hat{H}_k(x)$, and compute $\mathrm{D}_{W_1}(\hat{H}_k(x), H_k(x))$). We can now appreciate what is a plausible distance curve between two human-based processes, and with that, we can better discern that a particular system gives good but not perfect representation to human levels of production variability (for example, note the overlap between the two distributions in Figure 10.5). Upon visual inspection of analogous divergence distributions for different sampling strategies, we find similar shapes. In Figure 10.6, to present results for many decoding settings, tasks and probes, we

[7] This leads to a probability distribution whose pmf is hard if at all possible to characterise, meaning we cannot easily assess the probability of an outcome under the new distribution. But we have an explicit sampler for this new distribution, which is all our analysis tools require.

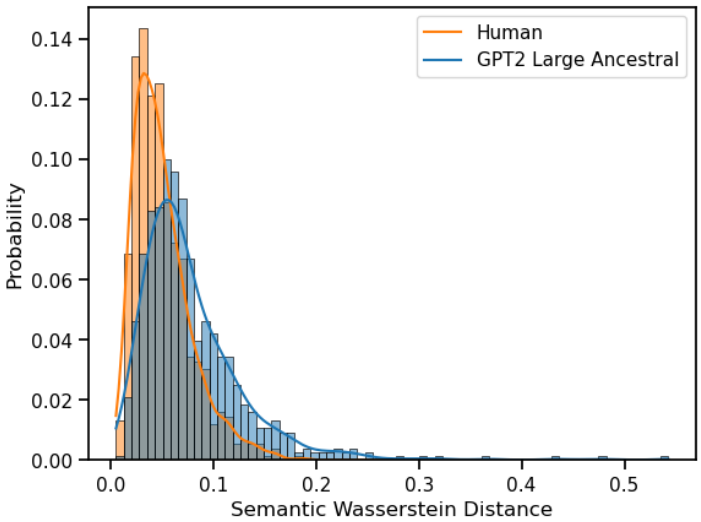


Figure 10.5: Distribution over Wasserstein distances $\mathrm{D}_{W_1}(C_{\mathrm{sem}}(x), H_{\mathrm{sem}}(x))$, in blue. Distances for a human control group, $\mathrm{D}_{W_1}(\hat{H}_{\mathrm{sem}}(x), H_{\mathrm{sem}}(x))$, in orange.

exploit this finding and summarise each divergence distribution using its mean. The leftmost red dots indicate the human control group.[8] We observe that two human groups agree more on the meaning of translations and simplifications than on their form, while for story generation the two groups agree more on surface form and basic structures and less on the semantic content of the stories.

Results. Figure 10.6 shows that most decoding settings are close to unbiased sampling, which in turn is in the same ballpark as the human control. This indicates that text generators capture the space of plausible human productions well when coupled with most decoding algorithms, though not as well as another human language process does. Decoding settings form many clusters, and for all tasks except open-domain dialogue, unbiased samples best match human variability. This suggests that, within the limits of decoding configurations typically considered as appropriate, different token-level decoding strategies often have a similar effect on a generator's ability to reproduce human production variability along our three probes. Altogether, these findings inform us about an often neglected aspect of decoding algorithms, namely their effect on the model's representation of uncertainty (rather than their ability to select individual high-quality generations).

[8] Except for dialogue as five references are too little to create a control group.

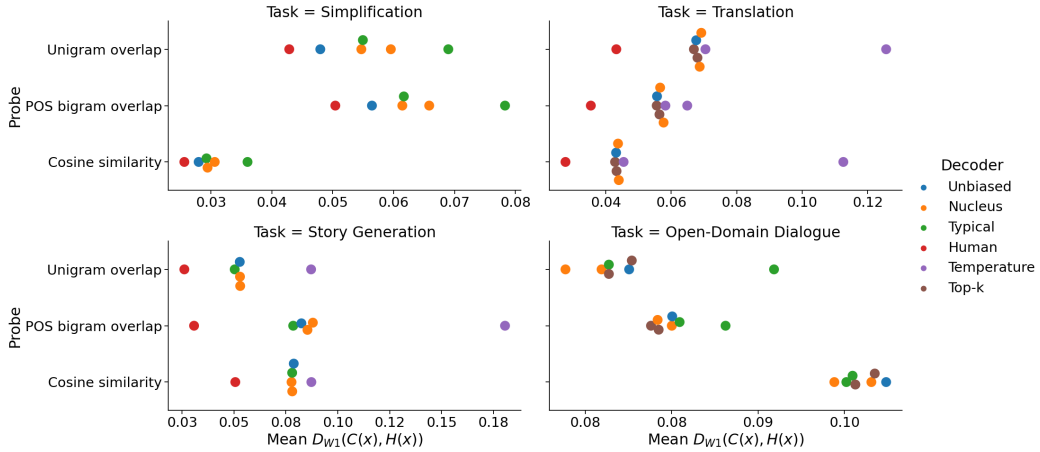


Figure 10.6: Mean Wasserstein distances $\mathrm{D}_{W_1}(C(x), H(x))$ for (tasks, probe, decoder) tuples. Base models for each task are described in Section 10.3. Colour refers to decoding algorithm with various parameter settings. Clusters suggest that decoders often have similar effect. No 'Human' data point for dialogue, where five references are too little to create a control group.

10.6 Qualitative instance-level analysis

We now qualitatively analyse individual inputs for which a generator's uncertainty is miscalibrated to human variability—as detected by D_{W_1}. For each task, we use up to 10 human productions (5 for dialogue) and 10 generations. As similarity metrics k, we use again unigram overlap (lexical), POS bigram overlap (syntactic), and cosine similarity (semantic). While it is not a replacement for more standard NLG evaluation procedures, we argue that this level of analysis is complementary and crucial to gain deeper understanding of a generator's representation of uncertainty.

Variability underestimation in translation and simplification. We have seen that in translation and simplification, generators' self-variability is lower than human variability (Section 10.5.1). We now zoom in on examples from these two tasks, inspecting instances that show inadequate fitness to human variability on all linguistic levels (i.e., $\mathrm{D}_{W_1}(M_k(x), H_k(x))$ is high for all k). The most severe cases of miscalibration for OpusMT are all instances of variability underestimation.[9] For most of these inputs, generations are virtually or completely identical, while a few present slightly higher but still substantially lower variability than human productions. For example, ten humans translated the phrase '*reacted cautiously*' in the English source sentence '*Several companies have thus far reacted*

[9]We select instances with $\mathrm{D}_{W_1} > 0.3$ for unigram overlap and $\mathrm{D}_{W_1} > 0.2$ for POS bigram and semantic overlap; we find 7 such instances. These thresholds are chosen based on distribution plots of instance-level distances (see, e.g., Figure 10.2b).

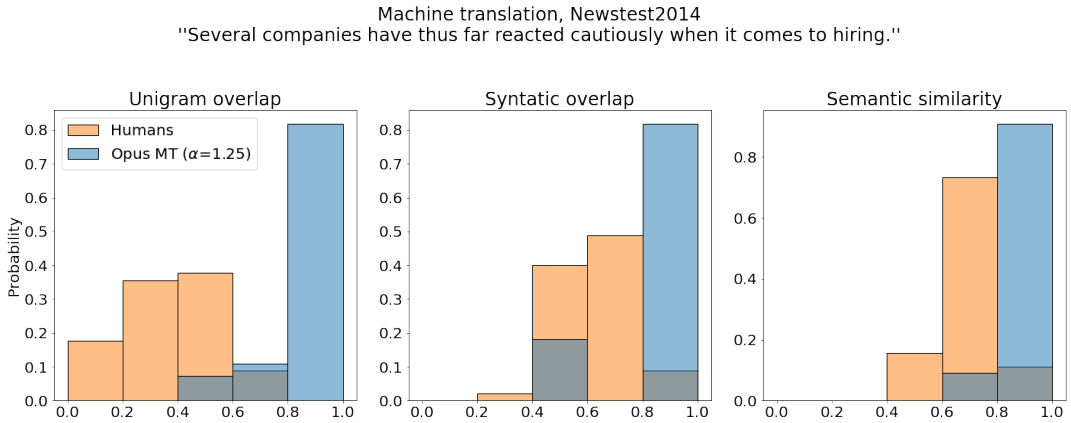


Figure 10.7: An example of variability underestimation in translation. Self-variability, $D_{W_1}(M_k(x), H_k(x))$, OpusMT with temperature sampling.

cautiously when it comes to hiring' in six different ways ('*zurückhaltend reagiert*', '*mit Vorsichtsmaßnahmen reagiert*', '*reagierten mit Zurückhaltung*', '*mit Vorsicht reagiert*', '*reagierten verhalten*') while all ten generated samples contain the German phrase *vorsichtig reagiert*', signalling that the generator's lexical rephrasing abilities do not generalise to this input instance (see Figure 10.7). For text simplification, we focus on instances where Flan-T5's uncertainty is not calibrated to human syntactic variability.[10] We observe that simplifications sampled from the generator are always syntactically more similar to each other than humans', indicating that the generator struggles to capture an important aspect of the text simplification task: that many (semantically equivalent) rewritings are possible for a text if the text's syntactic structure is altered.

Variability overestimation in dialogue. According to our estimates of human variability (Section 10.4), dialogue is the most open-ended task on all linguistic levels. We have hypothesised that this is due to the large variety of communicative act types plausible given any dialogue context. We have also seen that DialoGPT generally overestimates production variability (Section 10.5.1)—Figure 10.1 is one such example. Now we further inspect instances where cross-variability is miscalibrated with respect to human outputs.[11] We find that the generator's bad fit can be due to very short and generic responses (e.g., 'Well...', "haha", "Ahem", "Well done!"), but is more often due to the presence of fluent yet very diverse and often inadequate samples. For such instances, not only is the generator's cross-variability miscalibrated—self-variability, too, is overestimated on all linguistic levels. In particular, the generator's poor calibration to lexical and syntactic variability is related to its inability to choose the correct dialogue

[10] $\mathrm{D}_{W_1}(M_{\mathrm{syn}}(x), H_{\mathrm{syn}}(x)) > 0.2$; 49 instances.
[11] $\mathrm{D}_{W_1}(C_k(x), H_k(x)) > 0.2$ for all k; 15 instances.

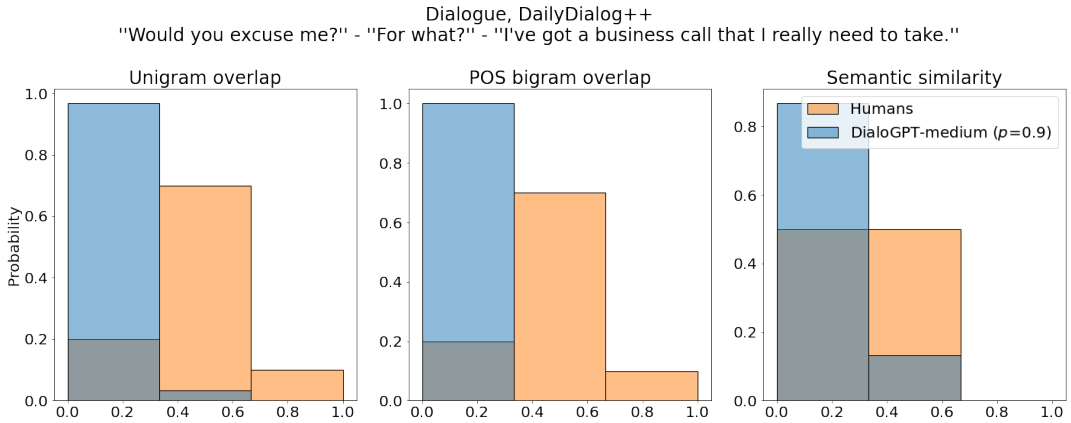


Figure 10.8: An example of variability overestimation in open-domain dialogue. Cross-variability, $D_{W_1}(C_k(x), H_k(x))$, DialoGPT Medium with nucleus sampling.

acts (or favouring an excessive variety of dialogue acts). In an example instance where the last dialogue turn goes 'I've got a business call that I really need to take' (see Figure 10.8), humans all reply with short affirmative responses ('Okay! Please.', 'Well! Go on.', 'Sure, why not!', 'Sure! Go ahead.', 'Yes! Sure.') while the model's responses are mostly lengthy statements, sometimes not particularly coherent ones (e.g., 'You don't need a business call. You need a friend').

Variability in lack of situational grounding. We have observed that human-written stories in the *WritingPrompts* dataset show lower variability than human dialogue responses, and hypothesised that this may be in part due to contextual pressures that constrain variability (Section 10.4). We now analyse instances flagged by our probe as cases of badly calibrated semantic cross-variability for GPT-2.[12] For one of these, the prompt refers to a context the model does not have access to ('all top level comments in this prompt take place in the same world, so make them all fit together'). Because they are conditioned on and reuse that context, human stories are quite similar to each other; generations, instead, show much lower pairwise similarity both when sampled jointly with the human productions and with themselves (see Figure 10.9). The lack of relevant situational grounding the model more uncertain than it should be for this instance.

10.7 Discussion and conclusions

Variability is an intrinsic property of human language production. If they are to be considered as good statistical models of human language production, text generators should exhibit plausible levels of variability. However, in NLG, the

[12] $\mathrm{D}_{W_1}(C_{\mathrm{sem}}(x), H_{\mathrm{sem}}(x)) > 0.3$; 7 instances.

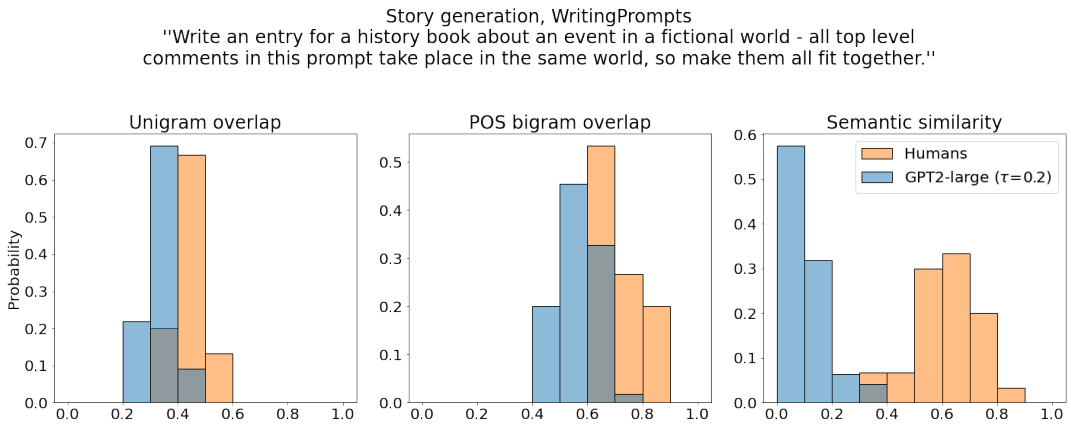


Figure 10.9: An example of variability overestimation in story generation. Cross-variability, $D_{W_1}(C_k(x), H_k(x))$, GPT-2 Large with typical sampling.

widespread practice is (i) collecting only one 'reference' production for each input and (ii) evaluating only a single generation. To appreciate the impact of this incongruity empirically, we analyse multiple-reference datasets for four NLG tasks, and show that each task has its own plausible levels of lexical, syntactic, and semantic variability. We connect production variability to aleatoric uncertainty, and evaluate neural text generators in terms of whether their representation of uncertainty is calibrated to the levels of variability observed in humans. We find, overall, that NLG systems are well calibrated to human levels of variability. Yet they slightly overestimate production variability in open-ended tasks and underestimate it in more constrained tasks. Moreover, we observe that most popular decoding algorithms all have a similar, limited effect on the generators' ability to reproduce human variability.

Statistical evaluation. We advocate for more widespread usage of instance-level probing of NLG systems as a way to evaluate their statistical fitness, not just along the dimensions we cover in this study but with respect to any other quality of interest. This approach contrasts with corpus-level analyses of NLG systems (e.g., Pillutla et al., 2021; Meister and Cotterell, 2021; Pimentel et al., 2022) and thanks to its greater interpretability, it builds trust in the ability of generators to reproduce human-like statistics when situated in specific linguistic contexts rather than 'globally', over a possibly heterogeneous corpus.

Impact of data collection. Our analysis relies on multiple-reference datasets, which are scarce for NLG tasks. Even though, for single-reference datasets, we cannot perform a similar instance-level analysis, this does not entail that our observations do not extend to such datasets—we might simply not have the data to expose them. The way in which multiple references are gathered may impact

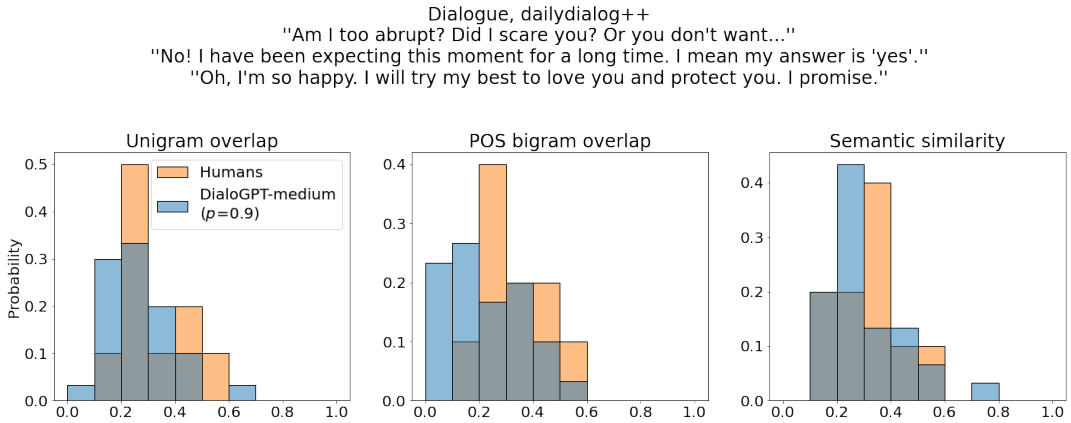


Figure 10.10: An example of well calibrated cross-variability in open-domain dialogue. $D_{W_1}(C_k(x), H_k(x))$, DialoGPT Medium with nucleus sampling.

the variability in productions. For example, asking a single annotator to produce several distinct references might artificially increase the diversity of responses. Conversely, asking several independent annotators might decrease diversity for they might resort to similar responses that come to mind quickly (or, in fact, the opposite might be true if they interpret the linguistic context differently). In this work, we do not distinguish between individual-level and population level variability, yet the analysis tools that we propose allow for such distinction.

Other quality dimensions. It is possible that a model fits various statistical properties of the human process (under $M_k(x)$, under $C_k(x)$, and for various choices of k) meanwhile none of its probable responses are deemed satisfactory as a whole by humans. This is why we shall think of our tools as (statistical) probes. We indeed find interesting instances that show good fit in terms of our similarity probes but whose outputs may be perceived as inadequate. Manual inspection reveals that a marriage proposal in one of the dialogues (Figure 10.10) is followed by a few incoherent model responses (e.g.., 'Thank you. It's not a question of the strength or weakness of the plot. I think it all falls within my capacity.'), some dispreferred ones ('If you want to have a hug?'; see Levinson, 1983), and some with negative affect ('I don't need your love. I know where you are coming from and I trust you will do the same.'). Exhaustively defining all aspects of perceived quality (or human-likeness) is a strenuous endeavour which is highly dependent on the use case of the generation system. Our similarity probes can be replaced with quality metrics which capture aspects (e.g., affective content, toxicity, or readability) that are considered relevant for any given application.

Outlook. In the future, we plan to devise new ways of improving the calibration of models' uncertainty (Zhao et al., 2022; Zhang et al., 2022b; Lee et al.,

2022), e.g., steering generators with sequence-level decoding algorithms (Eikema and Aziz, 2022a), and to investigate the relation between uncertainty and perceived generation quality (e.g., Kuhn et al., 2023): while we use human levels of variability as a target, desirable levels of variability may deviate from human statistics for specific applications. Future work should also study uncertainty as a function of a more complex notion of discourse context (as we did in Chapter 7) as well as attempt to disentangle uncertainty over communicative goals and realisations (Stasaski and Hearst, 2023). This is an important avenue not only toward more practically useful generators but also toward using NLG systems as reliable computational models of language production.

Chapter 11

Measuring utterance predictability with neural text generators

The content of this chapter is based on the following paper, which is under submission at the time of writing:

> Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. Information Value: Measuring Utterance Predictability as Distance from Plausible Alternatives. To appear in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Republic of Singapore. Association for Computational Linguistics.

Mario produced the research idea. Mario performed the majority of the experiments, to which Sarenne also contributed. Mario and Sarenne performed the analysis and wrote the article. Raquel provided advice throughout the project, she reviewed and revised the manuscript. The text in this chapter overlaps with that of the original publication.

11.1 Introduction

Measuring the amount of information carried by a linguistic signal is fundamental to the understanding and computational modelling of human language processing. Measures of information are used in psycholinguistic and neurobiological models of language processing (Levy, 2008a; Willems et al., 2016; Futrell and Levy, 2017; Armeni et al., 2017), to study the processing mechanisms of neural language models (Futrell et al., 2019; Davis and van Schijndel, 2020; Sinclair et al., 2022), and as a learning and evaluation criterion for language modelling (under the guise of 'perplexity'). As discussed in Section 9.2 of the background chapter, the amount of information carried by a linguistic signal is intrinsically related to its predictability (Hale, 2001; Genzel and Charniak, 2002), and this connection is summarised by surprisal (Shannon, 1948), perhaps the most widely used measure of information. Predictable signals carry low amounts of information—i.e., surprisal—as they are already expected to occur given the context in which they are produced. Conversely, unexpected signals carry higher surprisal.

Proper estimation of the surprisal of an utterance would require computing probabilities over a high-dimensional, structured, and ultimately unbounded event space. It is thus common to resort to chaining token-level surprisal estimates, nowadays typically obtained from neural language models. However, token-level autoregressive approximations of utterance probability have a few problematic properties, which we have already discussed in the introduction to Part 3. First, different realisations of the same concept or communicative intent compete for probability mass (Holtzman et al., 2021). Moreover, token-level surprisal estimates conflate different dimensions of predictability, which makes it difficult to appreciate whether the information carried by an utterance is a function, for example, of the unexpectedness of its lexical material, syntactic arrangements, semantic content, or speech act type (Arehalli et al., 2022; Kuhn et al., 2023).

We propose an intuitive characterisation of the information carried by utterances, *information value*, which is unaffected by these issues. It measures predictability over the space of full utterances and it explicitly accounts for multiple dimensions of predictability (e.g., lexical, syntactic, and semantic). Given a linguistic context, the information value of an utterance is a function of its distance from the set of contextually expected alternatives. The intuition is that if an utterance differs largely from alternative productions, it is an unexpected contribution to discourse with high information value. We obtain empirical estimates of information value by sampling alternatives from neural text generators and measuring their distance from a target utterance using interpretable distance metrics. Estimates are evaluated in terms of their ability to predict and explain human reading times and acceptability judgements in dialogue and text.

We find information value to have stronger psychometric predictive power than aggregates of token-level surprisal for acceptability judgements in spoken and

written dialogue, and to be complementary to surprisal as a predictor of reading times. Furthermore, our suite of interpretable information measures provides insights into the processing mechanisms underlying comprehension behaviour. It reveals, for example, that utterance acceptability in dialogue is largely determined by semantic predictability while reading times are more affected by lexical and syntactic expectations.

Beyond being a powerful tool for the analysis of comprehension behaviour (Meister et al., 2021; Shain et al., 2022; Wallbridge et al., 2023), information value can be used to model audience-aware language production strategies (Genzel and Charniak, 2002; Doyle and Frank, 2015a; Xu and Reitter, 2018; Verma et al., 2023) and to design mechanisms that reproduce them in natural language generation systems (Wei et al., 2021; Meister et al., 2023).

11.2 Alternatives in semantics and pragmatics

The measure of predictability presented in this chapter takes inspiration from the concept of alternatives in semantics and pragmatics (Stalnaker, 1978; Gazdar, 1979; Horn, 1972; Grice, 1975; Rooth, 1996; Levinson et al., 2000). In language production and comprehension, humans constantly process information about other things they could say or that could have been said. Reasoning about alternatives has been argued to be at the basis of the use of questions (Hamblin, 1976; Groenendijk and Stokhof, 1984; Ciardelli et al., 2018), focus (Rooth, 1992; Wagner et al., 2005; Beaver and Clark, 2009), and implicatures (Carston, 1998; Degen and Tanenhaus, 2015, 2016; Zhang et al., 2023).

Recently, alternative sets generated with the aid of language models have been used to provide empirical evidence that pragmatic inferences of scalar implicature depend on listeners' context-driven uncertainty over alternatives (Hu et al., 2022, 2023). Hu et al. (2022) generate sets of plausible words in context, within scalar constructions, then embed and cluster the resulting sentences to simulate conceptual alternatives (Buccola et al., 2022). Reasoning over word- and concept-level alternatives is operationalised through surprisal and entropy. To our knowledge, ours is the first study to use language models for the generation of full utterance-level alternatives.

11.3 Alternative-based information value

Given a context x, a speaker may produce a number of plausible utterances. We refer to these as A_x, the *alternative set*. We define the *information value* of an utterance y in a context x as the distribution of distances of y from the set of alternative productions A_x that are expected given x, measured with a distance metric d:

$$I(Y = y | X = x) \coloneqq d(y, A_x) \tag{11.1}$$

This distribution characterises the predictability of y in its context. Large distances indicate that y differs substantially from expected utterances, and thus that y is a surprising next utterance.

11.3.1 Computing information value

In Equation 11.1, we define information value as an abstract notion of the unpredictability, or unexpectedness of an utterance. In practice, computing the information value of an utterance requires a method to generate alternative sets $|A_x|$, a metric with which to measure the distance of an utterance from its alternatives, and a means with which to summarise distributions of pairwise distances.

Generating alternative sets. Since the 'true' alternative sets entertained by a human comprehender are not attainable, we propose generating alternative sets algorithmically, via neural text generators. Furthermore, we compare utterances and alternatives with a selection of distance metrics and summary statistics which enable an exploration into the representational form of alternatives in human language processing (Gotzner and Romoli, 2022). The plausibility, or human-likeness of the generations is clearly another important factor. Our approach builds on findings from the previous chapter, where we observed that the predictive distribution of neural text generators is well aligned to human variability (as measured with the same distance metrics we use here): while not all generations are guaranteed to be of high quality, their low-dimensional statistical properties match those of human productions. This should allow us to obtain faithful distance distributions $d(y, A_x)$ and thus accurate estimates of information value.

Measuring distance from alternatives. We quantify the distance of a target utterance from an alternative production using the three interpretable distance metrics introduced in Chapter 10. They are reported again here for convenience. **Lexical:** Fraction of common n-grams in two utterances, with $n \in [1, 2, 3]$ (i.e., number of matching n-gram occurrences divided by the total number of n-grams in both strings). **Syntactic:** Fraction $\text{syn}_n(y', y)$ of common part-of-speech (POS) n-grams in two utterances. **Semantic:** Cosine and euclidean similarity $\text{sem}(y', y)$ between the sentence embeddings of two utterances (Reimers and Gurevych, 2019). These distance metrics characterise alternative sets at lexical, structural, and conceptual levels (Katzir, 2007; Fox and Katzir, 2011; Buccola et al., 2022).

Summarising distance distributions. Information value is a distribution over distances (Equation 11.1). To summarise this distribution, we explore *mean* as the expected distance (under a uniform distribution over alternatives) or as the distance from a prototypical alternative, and *min* as the distance of y from the

closest production, implicating that proximity to a single alternative is sufficient to determine predictability.

11.4 Experimental setup

11.4.1 Language models

We generate alternative sets using neural autoregressive language models (LMs). For the dialogue corpora, we use GPT-2 (Radford et al., 2019), DialoGPT (Zhang et al., 2020c), and GPT-Neo (Black et al., 2021). For the text corpora, we use GPT-2, GPT-Neo, and OPT (Zhang et al., 2022a). The text models are pre-trained, while dialogue models are fine-tuned on the respective datasets. Further details on fine-tuning and perplexity scores are in Appendix C.2.

Generating alternatives. To generate an alternative set A_x, we sample from $p_{LM}(Y|X=x)$. We experiment with four popular sampling strategies to ensure that the quality of our surprise estimates is not dependent on a particular strategy—or, if it is, that we are not overlooking it. We select (i) *unbiased* (ancestral or forward) sampling (Bishop, 2006; Koller and Friedman, 2009), (ii) *temperature sampling* ($\alpha \in [0.75, 1.25]$), (iii) *nucleus sampling* (Holtzman et al., 2019) ($p \in [0.8, 0.85, 0.9, 0.95]$), and (iv) *locally typical sampling* (Meister et al., 2023) ($\tau \in [0.2, 0.3, 0.85, 0.95]$). We post-process alternatives to ensure that each contains only a single utterance.

11.4.2 Psychometric data

Using five corpora, we study two main types of psychometric variables that rely on different underlying processing mechanisms (Gibson and Thomas, 1999; Hofmeister et al., 2014): acceptability judgements and reading times.

Stimuli for **acceptability judgements** typically consist of isolated sentences that are manipulated automatically or by hand to assess a *grammatical* notion of acceptability (Lau et al., 2017; Warstadt et al., 2019). The effect of context on acceptability is still relatively underexplored, yet contextualised judgements arguably capture a more natural, intuitive notion of acceptability. In this study, we use some of the few datasets of in-context acceptability judgements which examine grammaticality as well as semantic and pragmatic plausibility.

Previous literature regarding the predictive power of language models for reading behaviour has focused on the relationship between per-word surprisal and **reading times** (Keller, 2004; Wilcox et al., 2020; Shain et al., 2022; Oh and Schuler, 2022). We define utterance-level reading time as the total time spent reading the constituent words of the utterance, following previous studies of utterance-level surprisal (Meister et al., 2021; Amenta et al., 2022).

Switchboard and DailyDialog. Participants were presented with a short contextual sequence of dialogue turns followed by a potential upcoming turn, and asked to rate its plausibility on a scale from 1 to 5. Judgements were collected by Wallbridge et al. (2022) for (transcribed) spoken dialogue, from the Switchboard Telephone Corpus (Godfrey et al., 1992), and for written dialogue, from DailyDialog (Li et al., 2017). For each corpus, 100 items were annotated by 3-6 participants. Annotation items consist of 10 dialogue contexts, each followed by the true next turn and by 9 turns randomly sampled from the respective corpus.[1]

Clasp. Participants were presented with sentences from the English Wikipedia in and out of their document context and asked to judge acceptability using a 4-point scale (Bernardy et al., 2018). The original sentences are round-trip translated into 4 languages (Czech, Spanish, German and French) to obtain varying degrees of acceptability; the context is not modified. This dataset contains 500 stimuli, annotated by 20 participants.[2]

Provo. This corpus consists of 136 sentences (55 paragraphs) of English text from a variety of genres, including online news articles, popular science, and fiction. Eye movement data was collected from 84 native American English speakers (Luke and Christianson, 2018). We use the sum of word-level reading times (IA-DWELL-TIME, the total duration of all fixations on the target word) of constituent words to obtain utterance-level measures (Meister et al., 2021).

Brown. This corpus consists of self-paced moving-window reading times for 450 sentences (12 passages) from the Brown corpus of American English. Reading times were collected from 35 native English speakers (Smith and Levy, 2013).

11.5 The psychometric predictive power of information value

We evaluate our empirical estimates of information value in terms of their psychometric predictive power: can they predict comprehension behaviour recorded as human acceptability judgements and reading times? We test the robustness of this predictive power and compare it to previously proposed utterance-level surprisal aggregates including mean, variance, and a range of summation strategies; see Appendix C.4 for full definitions.

[1] Negative turns were sampled to span the range of conditional surprisals expected from true dialogue continuations (see Section 3.4 in Wallbridge et al., 2022). Acceptability judgements are available at `https://data.cstr.ed.ac.uk/sarenne/INTERSPEECH2022/`.

[2] We only use judgements collected in context: `https://github.com/GU-CLASP/BLL2018`.

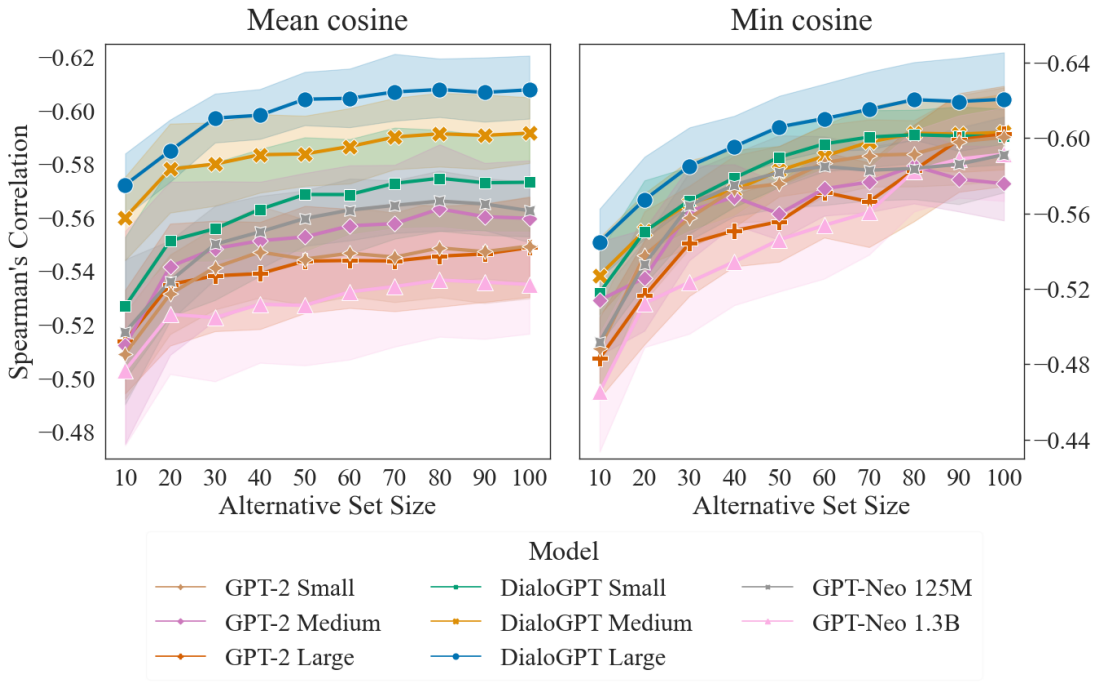


Figure 11.1: Spearman correlation between semantic information value and mean acceptability judgements in SWITCHBOARD. Confidence intervals display variability over 11 sampling strategies.

For each corpus in Section 11.4.2, we measure the correlation between information value and the respective psychometric variable, which is the average in-context acceptability judgement for DAILYDIALOG, SWITCHBOARD, and CLASP, and the total utterance reading time normalised by utterance length for PROVO and BROWN.[3] Alternative sets are generated using the language models and sampling strategies described in Section 11.4.1. Lexical, syntactic, and semantic distances are computed in terms of the distance metrics presented in Section 11.3.1, for alternative sets of varying size ($[10, 20, ..., 100]$). The distributions of similarities in Equation 11.1 are summarised using *mean* and *min*, thus yielding scalar estimates of information value.

11.5.1 Predictive power

We obtain moderate to strong Spearman correlations between information value and psychometric data across all corpora with the models and sampling strategies from Section 11.4.1. Notably, estimates of semantic information value cor-

[3]We normalise by utterance length as it is an obvious correlate of total reading time and would have confounding effects on this analysis. In Section 11.6, we confirm our findings using mixed effect models that include utterance length as a predictor and total unnormalised reading time as a response variable.

	Information value	**Surprisal**
Acceptability *($x \propto y^{-1}$)*		
SWITCHBOARD	-0.702 *(semantic)*	-0.506 *(superlinear, $k=4$)*
DAILYDIALOG	-0.584 *(semantic)*	-0.457 *(superlinear, $k=2.5$)*
CLASP	-0.234 *(syntactic)*	-0.559 *(mean)*
Reading times *($x \propto y$)*		
PROVO	0.421 *(syntactic)*	0.495 *(variance)*
BROWN	0.223 *(lexical)*	0.220 *(mean)*

Table 11.1: Correlations of the best variants of information value and surprisal (in parenthesis) with psychometric data: mean acceptability judgements and length-normalised reading times.

relate with acceptability judgements at strengths approximately between -0.4 and -0.7 for SWITCHBOARD and between -0.3 and -0.6 for DAILYDIALOG (see Figure 11.1 for SWITCHBOARDand Appendix C.3 for all datasets). Estimates obtained with the best information value estimators, shown in Table 11.1, yield substantially higher correlations with acceptability than the best token-level aggregates of utterance surprisal, both as computed in our experiments and as reported in prior work (Wallbridge et al., 2022, 2023). Reading times, on the other hand, are aggregates of word-level psychometric data points and should thus naturally be easier to capture with word-level measures of predictability. Nevertheless, our best information value estimates correlate with reading times only slightly less strongly or comparably to surprisal; and additionally, they give us indications about the dimensions of unexpectedness that mostly affect reading behaviour (indeed these are mostly based on lexical and syntactic distances).

Overall, beyond building trust in our information value estimators, this evaluation demonstrates the benefit of their interpretability. The predictive power for lexical, syntactic, and semantic distances varies widely between corpora. Semantic distances are much more predictive for dialogue datasets than lexical or syntactic distances, while the inverse is true for the reading times datasets. We explore differences between the underlying perceptual processes employed for these two comprehension tasks further in Section 11.6.

11.5.2 Robustness to estimator parameters

We now study the extent to which our estimates are affected by variation in three main factors that determine information value: the alternative set size ($[10, 20, ..., 100]$), the language model, and the sampling strategy. We find a slight positive, asymptotic relationship between correlations and alternative set size for semantic information value in the dialogue corpora—information value estimates

Corpus	Level	Metric	Summary	N	Language Model	Sampling	ρ
	Lexical	Bigram	Min	70	DialoGPT Medium	Temperature 1.25	-0.436*
Switchboard	Syntactic	POS bigram	Min	100	DialoGPT Small	Ancestral	-0.440*
	Semantic	Cosine	Min	100	DialoGPT Large	Temperature 1.25	**-0.702***
	Lexical	Unigram	Min	80	DialoGPT Small	Ancestral	-0.383*
DailyDialog	Syntactic	POS trigram	Min	90	DialoGPT Large	Temperature 1.25	-0.359*
	Semantic	Cosine	Min	100	GPT-2 Large	Nucleus 0.9	**-0.584***
	Lexical	Trigram	Min	90	GPT-2 Large	Temperature 1.25	-0.210*
Clasp	Syntactic	POS Bigram	Min	100	GPT-2 Large	Nucleus 0.95	**-0.234***
	Semantic	Cosine	Min	90	OPT 1.3B	Temperature 0.75	-0.221*
	Lexical	Unigram	Min	10	OPT 125M	Typical 0.3	0.379*
Provo	Syntactic	POS Trigram	Min	10	GPT-2 Small	Nucleus 0.95	**0.421***
	Semantic	Euclidean	Min	100	OPT 125M	Nucleus 0.95	0.181
	Lexical	Bigram	Min	90	GPT-2 Small	Typical 0.3	**0.223***
Brown	Syntactic	POS Trigram	Mean	10	GPT-2 Medium	Typical 0.3	0.185*
	Semantic	Cosine	Min	100	GPT-Neo 125M	Nucleus 0.95	0.048

Table 11.2: Best information value estimator per corpus and metric. Spearman rank-correlation coefficients ρ, statistical significance ($p < 0.001$) is marked with a star. The highest correlations per dataset are in **bold**; the estimators (a combination of set size N, model, and sampling strategy) that generate them are taken as the 'best estimators' for that corpus and are used in Sections 11.6 and 11.7.

become more predictive as alternative set size increases (see, e.g., Figure 11.1). Set size does not significantly affect correlations for the reading times corpora. Moreover, while we do observe differences between models, and larger models tend to obtain higher correlations with psychometric variables, these results are not consistent across corpora and distance metrics (see Figures C.6 and C.7 in Appendix C.3). In light of recent findings regarding the *inverse* relationship between language model size and the predictive power of surprisal (Shain et al., 2022; Oh and Schuler, 2022), we consider it an encouraging result that the predictive power of information value does not decrease with the number of model parameters.[4] We do not observe a significant impact of decoding strategy on predictive power, regardless of alternative set size, as indicated by the confidence intervals in Figure 11.1.

In sum, estimates of information value do not display much sensitivity to alternative set generation parameters.[5] Therefore, for each corpus, we select the estimator (a combination of model, sampling algorithm, and alternative set size) that yields the best Spearman correlation with the psychometric data (Table 11.2). We use these estimators throughout the rest of this study.

[4]It remains to be seen whether this trend extends to larger language models, for which we lack computational resources.

[5]We obtain similar evidence of robustness to parameter settings using an intrinsic evaluation, reported in Appendix C.5.

11.6 In-depth analysis of psychometric data

Using information value, we now study which dimensions of predictability effectively explain psychometric data. This allows us to qualitatively analyse the processes humans employ while reading and to assess acceptability. We also define two additional measures derived from information value (Section 11.6.1) and use them as explanatory variables in linear mixed effect models to predict per-subject psychometric data.[6] For the dialogue corpora and CLASP, our mixed effect models predict in-context acceptability judgements. For the reading times corpora, our models predict the total time spent by a subject reading a sentence, as recorded in self-paced reading and eye-tracking studies. This is the sum, over a sentence, of word-level reading times (more details in Appendix C.7). We include random intercepts for (context, target) pairs in all models.

Analysis procedure. For each corpus, we first test models that include a single predictor beyond the baseline: i.e., information value measured with all distance metrics and both *mean* and *min* as summary statistics (see Section 11.3.1). Based on the fit of these single-predictor models, we select the best lexical, syntactic, and semantic distance metrics (with the corresponding summary statistics) and use them to instantiate three-predictor models for each of the derived measures of information value. Following Wilcox et al. (2020), we evaluate each model relative to a baseline model which includes only control variables. Control variables are selected building on previous work (Meister et al., 2021): solely an intercept term as a baseline for acceptability judgements and the number of fixated words for reading times (more details in Appendix C.7). As an indicator of explanatory power, we report ΔLogLik, the difference in log-likelihood between a model and the baseline: a positive ΔLogLik value indicates that the psychometric variable is more probable under the comparison model. We also report fixed effect coefficients and their statistical significance. The full results are shown in Table 11.3.

11.6.1 Derived measures of information value

Inspired by information-theoretic concepts used in previous work to study the predictability of utterances (e.g., Genzel and Charniak, 2002; Keller, 2004; Xu and Reitter, 2018), as well as in the previous chapters (Part 2), we define two derived measures of information value and assess their explanatory power.

Out-of-context information value is the distance of an utterance y with respect to the set of alternative productions A_x expected given the empty context ϵ:

$$I(Y=y) \coloneqq I(Y=y|X=\epsilon) \tag{11.2}$$

[6] Three more derived measures are defined in Appendix C.6. We found them to be less predictive than those presented here.

Summ.	Level	Metric	Switchboard		DailyDialog		Clasp		Provo		Brown	
			β	ΔLogLik	β	ΔLogLik	β	ΔLogLik	β	ΔLogLik	β	ΔLogLik
Mean	Lexical	Unigram	-0.273	1.874	-0.683	2.152	0.594	0.206	2.309	9.967	2.409	9.679
		Bigram	-1.35	4.687	-2.761*	7.179	-1.573	2.717	1.976	10.971	1.622	9.076
		Trigram	-2.401	**8.315**	-1.843	7.089	-2.514	**5.857**	1.982	**12.169**	1.891	11.974
	Syntactic	POS Unigram	0.204	0.605	3.399**	**6.707**	0.914	0.106	1.902	8.413	0.958	6.627
		POS Bigram	0.398	1.366	1.835	3.147	-0.648	-0.073	3.813**	13.861	1.331	7.291
		POS Trigram	-0.159	**2.488**	0.767	2.505	-2.011	2.274	5.527**	**21.798**	1.475	8.194
	Semantic	Cosine	-8.664**	29.034	-6.988**	21.207	-1.235	0.030	0.237	6.661	0.714	6.633
		Euclidean	-8.665**	29.263	-7.11**	21.994	-1.535	0.617	0.221	**6.864**	0.766	**6.833**
Min	Lexical	Unigram	-3.927**	7.701	-4.454**	10.244	-0.866	0.005	2.219	9.649	2.39	9.059
		Bigram	-1.017	1.629	-4.614**	**10.876**	-1.937	1.639	1.882	9.490	2.121	8.426
		Trigram	-1.774	3.396	-1.969	3.311	-2.757*	3.714	1.997	10.337	3.689**	**13.913**
	Syntactic	POS Unigram	-0.52	0.927	0.356	1.985	-2.931*	4.915	5.45**	21.187	1.947	**8.633**
		POS Bigram	1.052	0.901	-2.933*	5.067	-5.356**	**13.539**	4.494**	16.292	1.404	6.854
		POS Trigram	0.758	0.993	-3.26*	5.732	-3.104*	4.124	4.956**	18.394	1.362	6.706
	Semantic	Cosine	-9.888**	**34.204**	-9.01**	**30.408**	-1.982	0.979	0.548	6.476	0.661	6.164
		Euclidean	-7.696**	23.375	-8.901**	29.868	-2.501	**2.094**	0.507	6.567	0.699	6.020

Table 11.3: Results of single-predictor linear mixed effect models: fixed effect coefficients β and ΔLogLik. Statistical significance of fixed effects is marked with one ($p < 0.01$) or two stars ($p < 0.001$). Information value estimates are obtained according to Equation 11.1. For each corpus and each level (lexical, syntactic, and semantic), the best ΔLogLik is marked in **bold**. These are the estimators we use whenever we talk about 'best predictors' in this chapter.

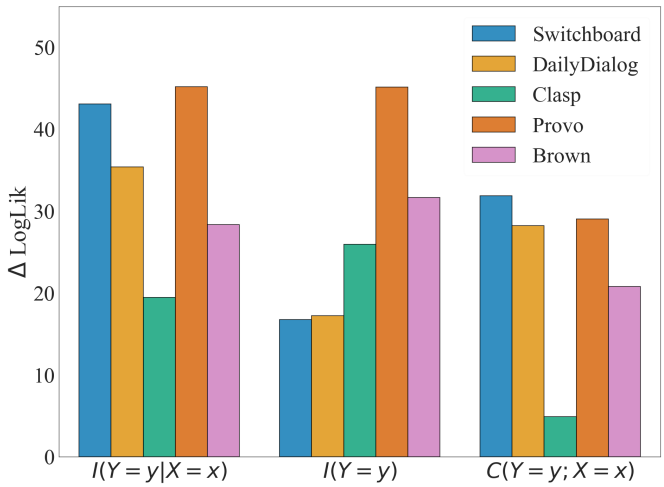


Figure 11.2: Explanatory power of information value and its derived measures, as defined in Section 11.6.1.

It captures the plausibility of y regardless of its context of occurrence.

Context informativeness is the reduction in information value for y contributed by context x:

$$C(Y=y;X=x) \coloneqq I(Y=y) - I(Y=y|X=x) \tag{11.3}$$

This quantifies the extent to which a context restricts the space of plausible productions in such a way that y becomes more predictable.

11.6.2 Acceptability judgements

We generally expect an inverse relationship between information value and in-context acceptability judgements: information value is lower when a target utterance is closer to the set of alternatives a comprehender may expect in a given context. Furthermore, we expect grammaticality and semantic plausibility—two factors known to affect acceptability (Sorace and Keller, 2005; Lau et al., 2017)—to play different roles in dialogue and text. For the dialogue corpora, we expect semantic-level variables to have high explanatory power, as they can identify utterances with incoherent content and implausible underlying dialogue acts. Lexical and syntactic information value may be more explanatory of acceptability in CLASP, where stimuli are generated via round-trip translation and thus may contain disfluent or ungrammatical sentences (Somers, 2005).

SWITCHBOARD and DAILYDIALOG. For both dialogue corpora, semantic information content is by far the most predictive variable (Table 11.3), especially when *min* is used as a summary statistic. Responses to the same dialogue context can exhibit great variability and being close to a single expected alternative—in terms of semantic content and dialogue act type—appears to be sufficient for an utterance to be considered acceptable. Our analysis of derived measures (Figure 11.2) further indicates that acceptability is mostly determined by the in-context predictability of an utterance. The high explanatory power of context informativeness (almost twice that of out-of-context information value) suggests that contextual cues override inherent isolated plausibility.

CLASP. Syntactic information value is the best explanatory variable for acceptability judgements in CLASP (Table 11.3). This suggests that comprehenders entertain expectations over non-lexicalised constructions (here, in the form of POS sequences)—a result which could complement findings on the processing of lexicalised constructions in reading (e.g., Tremblay et al., 2011) and eye-tracking studies (e.g., Underwood et al., 2004). In contrast to the dialogue corpora, estimates of in-context information value are less predictive than their out-of-context counterparts (Figure 11.2), which may be due to the previously discussed artificial nature of the CLASP negative samples. In sum, our results indicate that the

acceptability judgements in the CLASP corpus, even if collected in context, are mostly determined by the presence of startling surface forms than by semantic expectations.

11.6.3 Reading times

When reading, humans continually update their expectations about how the discourse might evolve (Hale, 2001; Levy, 2008a; Yan and Jaeger, 2020). This is reflected, for example, in the faster processing of more expected words and syntactic structures (Demberg and Keller, 2008; Smith and Levy, 2013). High predictive power for lexical and syntactic information value would support these findings. However, comprehenders also reason about semantic alternatives, e.g., to compute scalar inferences (Van Tiel et al., 2014; Hu et al., 2023). Our interpretable measures of information value help clarify the contribution of different types of expectations.

PROVO and BROWN. Syntactic information value is a strong predictor of eye-tracked reading times in PROVO, while lexical information value (in particular, based on trigram distances) is the only significant explanatory variable for the self-paced reading times in BROWN (Table 11.3), and only weakly so. Expectations over full semantic alternatives have a limited effect on reading times in both corpora, suggesting anticipatory processing mechanisms operate at lower linguistic levels. For both corpora, out-of-context estimates are at least as predictive as in-context estimates and higher than context informativeness (Figure 11.2), indicating that context modulation has a moderate impact on the negative effects of unusual syntactic arrangements and lexicalised constructions on reading speed.

11.7 Relation to utterance surprisal

We have shown alternative-based information value to be a powerful predictor for contextualised acceptability judgements and reading times. In fact, information value is substantially more predictive of acceptability than utterance surprisal (Section 11.5). We conclude with a focused comparison between these measures, considering whether they are complementary and why they might diverge.

11.7.1 Complementarity

Differences in predictive power between information value and surprisal (see Table 11.1) may reflect variations between the dimensions of predictability captured by the two measures. To investigate this possibility, we use both measures jointly for psychometric predictions. We focus on the dialogue corpora and PROVO,

	Switchboard	DailyDialog	Provo
Surprisal	6.63	5.08	59.04
Information value			
Lexical	8.32	10.88	12.17
Syntactic	2.49	6.71	21.80
Semantic	34.20	30.41	6.86
All	43.11	**35.42**	45.19
Joint			
+ *Lexical*	14.08	10.23	72.60
+ *Syntactic*	9.77	8.05	75.70
+ *Semantic*	34.37	26.98	68.61
+ *All*	**44.11**	30.55	**93.08**

Table 11.4: ΔLogLik for surprisal, information-value, and joint models.

where we observed the highest explanatory power for information value (Section 11.6). For each corpus, we fit linear mixed effect models with control variables, using the best surprisal and information value predictors (one per linguistic level) in isolation and jointly as fixed effects. The results of this analysis are displayed in Table 11.4.

In isolation, information value is a better predictor for the dialogue corpora. Including lexical, syntactic, and semantic information value on top of the best surprisal predictor (*Joint*) improves model log-likelihood substantially. Separately including each linguistic level reveals that semantic distance is largely responsible for improved fit, suggesting that surprisal fails to capture expectations over high-level linguistic properties of utterances such as speech act type, which are crucial for modelling contextualised acceptability in dialogue.[7]

For Provo, surprisal is the best explanatory variable. However, including the best information value predictors further improves model fit by 58%, demonstrating the complementarity of the two measures in predicting reading times (Table 11.4). Separately adding information value predictors shows the strongest boost comes from syntactic factors, which are known to have higher weight in human anticipatory processing than in language models' (Arehalli et al., 2022).

Overall, combining predictive information value with surprisal yields better models for all tested corpora, indicating that these measures capture distinct and complementary dimensions of predictability.

[7]This is true in spite of the aggregation function used; here, we report maximum (Switchboard) and superlinear surprisal (DailyDialog), the aggregates with the highest ΔLogLik for the two corpora.

Dataset	Summary	Level	Metric	Context Condition		
				Congruent	*Empty*	*Incongruent*
Switchboard	Mean	Lexical	Trigram	**8.32**	5.55	7.18
	Mean	Syntactic	POS Trigram	2.49	**3.00**	2.65
	Min	Semantic	cosine	**34.20**	7.64	10.94
	Surprisal (in context, max)			**6.63**	2.56	3.12
DailyDialog	Min	Lexical	Bigram	**10.88**	3.16	1.42
	Mean	Syntactic	POS Unigram	6.71	**6.89**	6.16
	Min	Semantic	Cosine	**30.41**	1.43	2.90
	Surprisal (in context, superlinear $k=1.5$)			**5.08**	0.99	2.35
Provo	Mean	Lexical	Trigram	**12.97**	12.94	11.86
	Mean	Syntactic	POS Trigram	**25.86**	15.20	12.94
	Mean	Semantic	Euclidean	8.53	**10.88**	8.33
	Surprisal (in context, superlinear $k=0.5$)			35.75	37.88	**39.00**

Table 11.5: ΔLogLik of single-predictor models for information value and surprisal across context conditions.

11.7.2 Effects of discourse context

While context greatly influences language comprehension (e.g., Chen et al., 2023), little attention has been given to its impact on surprisal estimates. We examine whether the dissimilar predictability estimates of information value and surprisal stem from differences in their sensitivity to context, comparing how they behave under congruent, incongruent, and empty context conditions. In each condition, alternative sets and token-level surprisal are computed in the true context (*congruent*), a context randomly sampled from the respective corpus (*incongruent*), or with no conditioning (*empty*). We quantify effects on the best information value and surprisal predictors as ΔLogLik, using single-predictor models described in Section 11.6.

Table 11.5 displays results for Switchboard, DailyDialog, and Provo. Congruent context produces a substantial effect on the predictive power of semantic information value for both dialogue datasets; for DailyDialog, we see a 20-fold increase over the empty context condition. Surprisal shows a similar trend, though far less pronounced. Syntactic information value is the least affected by context modulations. Though surprisal is a powerful predictor for reading times in Provo, the incongruent and empty context conditions are *more* predictive than the true context. Perhaps most concerning is the fact that estimates in incongruent contexts are the most predictive. In contrast, the most predictive information value (syntactic) is significantly more predictive for congruent contexts. Interestingly, information value in the control conditions is not

uninformative, likely reflecting the inherent plausibility of utterances.

Both information value and utterance surprisal display sensitivity to context, however, the effects on surprisal are less predictable and perhaps even undesirable for certain psychometric variables.

11.8 Discussion and conclusions

Humans constantly monitor and anticipate the trajectory of communication. Their expectations over the upcoming communicative signal are influenced by factors spanning from the immediate linguistic context to their interpretation of the speaker's goals. These expectations, in turn, determine aspects of language comprehension, such as processing cost, and strategies of language production. We present *information value*, a measure which quantifies the predictability of an utterance relative to a set of plausible alternatives; and we introduce a method to obtain information value estimates via neural text generators. In contrast to utterance predictability obtained by aggregating token-level surprisal estimates, information value captures variability above the word level and considers the impact of more abstract communicative units like dialogue acts. We validate our measure by assessing its psychometric predictive power, its robustness to parameters involved in the generation of alternative sets, and its sensitivity to discourse context.

Explaining psychometric behaviour. Using interpretable information measures centred around information value, we investigate the underlying dimensions of uncertainty in human acceptability judgements and reading behaviour. We find that acceptability judgements factor in base rates of utterance acceptability (likely associated with grammaticality) but are predominantly driven by semantic expectations. In contrast, reading time is more influenced by the inherent plausibility of lexical items and part-of-speech sequences.

Relation to surprisal. We compare information value to aggregates of token-level surprisal, finding differences in the dimensions of predictability captured by each measure and their sensitivity to context. Information value is a stronger predictor of acceptability in written and spoken dialogue and is complementary to surprisal for predicting eye-tracked reading times.

Interpretability through custom distance metrics. Our framework for the estimation of utterance information value allows great flexibility. Modellers can experiment with a variety of alternative set generation procedures, distance metrics, and summary statistics. While our selection of distance metrics characterises the relation of an utterance to its alternative sets at multiple interpretable linguistic levels, there is a large space of metrics that we have not tested in this

study. Syntactic distances, for example, can be computed using metrics that capture structural differences between utterances in a more fine-grained manner (e.g., difference in syntactic tree depth); semantic distances can be computed with a more taxonomical approach (e.g., Fellbaum, 2010) or using NLI models to capture semantic equivalence (Kuhn et al., 2023); and distances between dialogue act types can be detected using dialogue act classifiers (Stasaski and Hearst, 2023). We chose metrics based on our prior work validating them as probes for the extraction of uncertainty estimates from neural text generators (Chapter 10), but we hope future work will explore this space more exhaustively.

Computational complexity. An aspect of our method for the estimation of information value that we have not highlighted in this chapter is its computational cost. Because it involves drawing multiple full utterance samples from language models, our method is clearly less efficient than traditional surprisal estimation, which requires only a single forward pass. While we have observed that the psychometric predictive power of information value reaches satisfactory levels even with relatively low numbers of alternatives and small language model architectures (see, e.g., Figure 11.1), designing more efficient methods for the estimation of information value is an important direction for future research.

Outlook. Our information value framework allows considerable flexibility in defining alternative set generation procedures, distance metrics, or summary statistics. We hope it will enable further investigation into the mechanisms involved in human language processing, and that it will serve as a basis for cognitively inspired and audience-aware strategies of utterance selection for natural language generation (Wei et al., 2021), or for the interpretation of existing decoding algorithm, which in some cases (Eikema and Aziz, 2020, 2022b) are already implicitly optimising neural estimates of information value.

Chapter 12

Towards human-like production strategies in natural language generation systems

The content of this chapter is based on the following publication:

> Mario Giulianelli. 2022. Towards Pragmatic Production Strategies for Natural Language Generation Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7978–7984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ideas and writing for this position paper were produced by Mario. The text in this chapter overlaps with that of the original publication.

This chapter collects insights from the rest of the thesis into a conceptual framework for efficient and communicatively effective—i.e., pragmatic—language production. I would like to thank Raquel Fernández for her invaluable contribution in shaping the perspective expressed in this chapter, in the corresponding paper, and—really—throughout the thesis.

12.1 Introduction

Novelists choose the right words to keep readers engaged and enthused, good journalists can convey facts clearly and convincingly, while poets may want to surprise the reader. Teachers adapt their explanations to the level of their students, and the language of parents changes with the proficiency of their children, with the same objects described first using simplified funny expressions ('moo moo') and then more informative and discriminative names ('cow', 'calf'). Using language to communicate successfully requires effort. On the side of the language producer, it is first of all effortful to come up with words that truthfully correspond to one's communicative intent. Then, words must be actually produced, e.g., said out loud or typed on a keyboard. At the same time, the producer has to take into consideration whether the comprehender—for whom, too, linguistic communication is costly—will be able to infer the original intent. Comprehenders make efforts to pay attention to the utterance they are being addressed with, to interpret it, and to infer their interlocutor's communicative intent. Fortunately, these efforts are often not in vain. They allow people to exchange knowledge, ideas, plans, and to achieve goals.

This chapter presents a conceptual framework for the computational modelling of utterance production in variably complex communicative scenarios, which relies on three main notions: **communicative goals**, production and comprehension **costs**, and **utility**. I define these notions formally and then, in two case studies, I provide suggestions for their operationalisation in classic NLG tasks. In sum, I advocate modelling humans as decision makers striving for efficient and effective communication, and argue that human-like linguistic behaviour emerges as a result of reasoning about goals, costs, and utility. Learning to navigate the complex decision space defined by these notions is still an open problem: this chapter discusses possible promising directions.

12.2 Doing things with words

Communication always comes with a goal: speakers use words to change the state of the world. This section gives a characterisation of communicative goals, discusses the types of effort (or costs) necessary to achieve goals, and describes the rewards associated with successful communication.

12.2.1 Communicative goal

What do speakers do with words? The *communicative goal* (or *communicative intent*) of a speaker can be formulated as a function of the current state of the world $w \in W$:

$$G_s\colon\ W \to W,\quad w \mapsto w^* \tag{12.1}$$

where w^* is the intended future state of the world. Speaker s and audience a are included in w as they can be both conceptualised, and there is evidence that they are processed (Brown-Schmidt et al., 2015), as parts of the state of the world. For communication to be successful, the audience must be able to reconstruct the original communicative goal: their *decoded* transformation of the world, $D_a : W \rightarrow W$, must be such that $D_a(w) \approx G_s(w)$.[1]

Communicative goals shape and constrain a speaker's production choices: different utterance types typically correspond to different goals. The communicative goal of a referring utterance ('The black and white cat'), for example, is a state of the world where the audience is able to identify an entity in context. The transformation D_a required to achieve w^* is a change of attention by the audience. Statements ('The Sun is a star') are typically used when the purpose of an interaction is pure information transmission—e.g., when giving a scientific talk. In this case, the communicative goal is a state of the world in which the audience holds new beliefs, the ones intended by the speaker. D_a is a transformation of the belief state of the audience, and the communicative goal is achieved when $D_a(w) \approx G_s(w) = w^*$. All utterance types—e.g., questions, directives, and performatives—can be seen as strategies to achieve communicative goals. The same utterance type, and even the same utterance, can fulfil different goals: a blatantly false statement ('It never rains in Amsterdam') can be used for comedic effect rather than for conveying facts. For simplicity, in the rest of this chapter, we describe utterances as having a single communicative goal. Often, however, different goals are associated with the same utterance at the same time: a teacher can use a question ('Are you sure this is the right answer?') to inform their student that their answer is incorrect, while showing a positive attitude towards them—thereby striving for both epistemic and social utility. Our framework naturally generalises over such cases; when multiple communicative goals are involved, states of the world can be designed accordingly. To account for epistemic and social utility, for example, states of the world can be defined to include the audience belief state as well as their emotional state.

12.2.2 Production costs

Given the current and the intended future state of the world, w and $w* = G_s(w)$, a speaker *encodes* the communicative goal $G_s(w)$ into a mental representation of the intended state of the world: $E_s(G_s(w)) = e$. To use a slightly different vocabulary, this is the speaker's *conception* of the intended environment state. The speaker then *realises* e as an utterance r which is presented to the audience: $R_s(e) = r$. Two types of cost are associated with the encoding and realisation processes. Because the encoding process is inevitably lossy—mental representations

[1] I sometimes refer to $D_a(w)$ and $G_s(w)$ as D_a and G_s, as in our formulation these functions are always applied to the current state of the world w.

are compressed representations of the real state of the world—the speaker makes an effort to reduce information loss; I refer to this as the **encoding cost** C^E. The cost associated with executing a bit of behaviour r meant to be perceived by the audience (e.g., speaking, writing, or typing) is the **realisation cost** C^R. Both costs affect the decision making process of speakers. In addition, the speaker is influenced by the expected comprehension costs of the audience.

12.2.3 Comprehension costs

The speaker's communicative goal $G_s(w)$ is not observable by comprehenders. Given a state of the world w and the speaker's behaviour r, comprehenders *process* r into a reconstruction of the original mental representation, $P_a(r) = e' \approx e$, from which they *decode* the speaker's communicative goal: $D_a(e') = w' \approx G_s(w)$. Two types of cost are associated with the comprehension of an utterance. Speaker and comprehender are different individuals and therefore have different ways of encoding communicative goals into messages (Connell and Lynott, 2014). In the absence of a perfect model of the speaker's encoding mechanism, reconstructing e is a lossy and effortful process; I denote the corresponding cost as **processing cost**, C^P. The second cost results from interpreting e' in context—i.e., decoding from e the state of the world intended by the speaker. In other words, this is the effort required to *ground* the message in the environment. I refer to it as the **decoding cost** C^D. It is important to note that although processing and decoding costs are on the side of comprehenders, speakers estimate them and take them into account when making production decisions.

12.2.4 Utility

In what ways is the decision making process of speakers affected by these costs? Speakers are thought to be driven by efficiency concerns (Zipf, 1949; Jaeger and Tily, 2011): they strive to minimise the collaborative effort required to achieve their communicative goals (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989). The speaker's utility U_s can thus be defined as being inversely proportional to the joint production and comprehension costs required for goal achievement ($D_a \approx G_s$). Production costs can be reduced directly by the speaker, by putting less cognitive and physical effort in encoding and realisation. Comprehension costs, instead, need to be first estimated via a mental model of the audience's comprehension system (including their conceptual knowledge, perceptual capacity, language proficiency, etc.). The ability to form such mental models is often referred to as Theory of Mind (Premack and Woodruff, 1978) and it is deemed a fundamental social-cognitive skill for language acquisition and language use (Tomasello, 2003).

Speaker's utility is not only defined in terms of costs; speakers profit from getting things done with their words. Thus U_s is also directly proportional to

the positive cognitive, physical, and social effects that derive from achieving the intended state of the world w^*. Because, in practice, interlocutors often approach but do not reach w^* exactly, U_s can be defined as a function of $D_a(w)$ and $G_s(w)$ that quantifies the difference in positive effects between true and intended states of the world.

12.3 Case study 1: Reference games

This section demonstrates how to use our framework to conceptualise a communication scenario that corresponds to a classic NLG task, referring expression generation (Reiter and Dale, 1997; Krahmer and van Deemter, 2012). I will also provide concrete examples of how to model the costs and utility described in Section 12.2.

In a reference game, the goal is for participants to produce descriptions that allow comprehenders to identify the correct referent out of a set of candidates. These games have been extensively used in psycholinguistics to study human strategies for effective reference (Krauss and Weinheimer, 1964; Brennan and Clark, 1996; Hawkins et al., 2020b). This case study is based on a visually grounded reference game with two participants, a speaker s and a listener a. The speaker produces referring utterances r such as 'a boy cutting a cake' and the listener needs to identify the target image i^* among a set of similar images V, the visual context (see, e.g., Shore et al., 2018; Haber et al., 2019). The initial state of the world is one where the speaker is aware of the target referent while the listener has no information about it. Such a state of the world can be expressed as $w = (V, p_s, p_a)$, i.e., in terms of the speaker and listener's probability distributions p_s and p_a over candidate images V before anything is uttered ($r = \epsilon$, the empty string):[2]

$$p_s(I|V) : p_s(I = i^*|V) = 1 \tag{12.2}$$

$$p_a(I|V, \epsilon) : p_a(I = i|V, \epsilon) = \frac{1}{|V|} \quad \forall i \in V \tag{12.3}$$

Note that p_s is never observable by a, and for this scenario to be realistic, p_a should also not be observable by s. The communicative goal G_s is a transformation of w into w^*, a state of the world in which a identifies i^* as the target referent:

$$G_s(w) = (V, p_s, p'_a) \text{ with} \tag{12.4}$$

$$p'_a(I = i^*|V, r) = 1 \tag{12.5}$$

How can the costs associated with reaching this state of the world using utterance r be estimated? A computer vision model may be used to encode the

[2]This setup corresponds to one-shot reference games. In multi-turn dialogues, w should also include the game history.

communicative goal $w^* = (V, p_s, p'_a)$ into a mental representation. This model receives as input the visual context V and information about the target image p_s and yields a mental (abstract) representation $e = E_s(w^*)$. If this is, e.g., a model that produces image segmentations, the encoding effort C^E can be quantified as the uncertainty of the model over its segmentation decisions, as the number of output image segments, or, if the segments form a scene graph, as a measure of the graph complexity. The encoding e may then be fed to an NLG model R_s which *realises* it into an utterance $r = R_s(e)$. The realisation cost C^R can be computed as the utterance length, the depth of the syntactic tree corresponding to the utterance, or as a function of the distribution of vocabulary ranks for the sampled utterance tokens.

Next, r is received by the listener, who *processes* it into a reconstruction of the original mental representation: $e' = P_a(r)$. This can be achieved using a neural language model, the processing cost C^P being calculated as the model's cumulative surprisal (the sum of the per-word information content). From e' the listener decodes a state of the world w'. The decoding system may be one that measures the similarity of e' to candidate image embeddings and outputs a probability distribution over V. The decoding cost C^D can be estimated as the entropy reduction with respect to the prior probability $p_a(I|V)$ (the information gain), or as the increase in the target image's probability. Communication is successful if $p'_a(I = i^*|V, r) = 1$ (see Equation 12.5); in practice the condition is often relaxed to:

$$i^* = \arg\max_{i \in V} p'_a(I = i|V, r) \tag{12.6}$$

In a simplified reference game where p_a is observable by s, the speaker's positive utility U_s can be simply modelled as $\log p'_a(i^*|V, r) - \log p_a(i^*|V)$. In a more realistic scenario, either the speaker entertains a mental model of p_a and uses it to compute utility, or the listener must in turn execute a bit of behaviour to communicate the state of p'_a, for example by selecting an image through a simple decision rule (e.g., $\arg\max p'_a$). U_s can then be modelled as a binary reward based on the listener's behaviour: 1 for a correct guess, 0 for an incorrect one. Recall that U_s is not only a function of positive cognitive effects. It is also inversely proportional to the costs C^E, C^R, C^P, and C^D.

12.4 Case study 2: Text summarisation

This second case study demonstrates the generality of our framework by applying it to text summarisation, a widely studied NLG (and NLU) task with a large range of practical applications. When people summarise a text, they produce a concise and meaning-preserving version of that text with the goal of conveying to the audience the text's most important ideas. In NLP, texts have been typically summarised either via extraction of their most significant sentences (Luhn, 1958; Edmundson, 1969) or by the generation of fewer, new sentences (DeJong, 1982;

Banko et al., 2000). Here, we look at the second case, often referred to as *abstractive summarisation*, where a summariser s produces an utterance r made up of one or multiple sentences to succinctly report the main content of a text t to an audience a. The initial state of the world is one where the summariser knows the content of t while the audience has no information about it.

Summaries can have multiple communicative goals (sometimes simultaneously) roughly corresponding to practical goals of NLP summarisation systems. For example, the communicative goal G_s of a summary can be a transformation of the state of the world into one in which a knows the general topic of t and is interested in reading t. This setup roughly corresponds to headline generation, a classic abstractive summarisation task. If the practical goal of the summary, instead, is to make the audience aware of the main facts reported in a text, the communicative goal G_s is a transformation of the state of the world into one in which those facts are part of a's knowledge. This is the goal, for example, of summaries of financial, legal, or medical reports.

Let us focus on this second case, for which I will provide examples of how to model communicative goals, costs, and utility. A hierarchical language model with explicit attention over multiple sentences can be used to encode the document into a mental representation e. The encoding cost C^E can be quantified as the entropy of the attention distribution—the rationale being that it is harder to condense the information in a document in which each sentence contains salient details. The encoding e may then be fed to a generation model R_s which realises it into an utterance $r = R_s(e)$ (one or multiple sentences). The realisation cost C^R can be computed as the utterance length or as a function of the predicted tokens' probabilities. The summary r is received by the audience, for example via a neural language model pretrained on summaries, which processes it into a reconstruction of the original mental representation: $e' = P_a(r)$. The processing cost C^P can be calculated as the model's cumulative surprisal. From e', the audience decodes a new state of the world, one where it can hopefully answer factual questions about the target document correctly. The decoding system can be a question answering model (which can be as simple as a table-lookup and as complex as a response generation model) and the decoding cost C^D can be estimated as the system's reduction in uncertainty in answering a set of questions designed to probe understanding of the main content of the document—formulated, e.g., as key-value queries or using natural language. The speaker's utility U_s can be modelled as the accuracy of the audience in answering questions about the content of the document.

12.5 Pragmatic production strategies

Language producers are thought to balance their own production costs and their audience's comprehension costs in a way that minimises joint collaborative effort (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989) while attempting

to gain utility from successful communication. Nevertheless, most modern NLG systems, whose aim is arguably to reproduce the communicative behaviour of human language users, do not take into consideration the costs and utility for which humans are constantly optimising. As a major example, GPT-3 (Brown et al., 2020a), one among the best foundation models available for NLG, conflates all costs into a single next-word probability value. To generate words from this model, typically, next-word probabilities are passed to a decoding algorithm such as beam search or nucleus sampling (Reddy, 1977; Holtzman et al., 2019). This algorithm can be seen as a way to search through the space of possible utterances by following a simple utility-maximising decision rule, with higher probability utterances having higher utility. Future work should investigate decision making rules that take into account production and comprehension costs more explicitly, connecting them to the goal of the linguistic interaction. The Rational Speech Act model (RSA; Frank and Goodman, 2012a) is a compelling solution: it was shown to optimise the trade-off between expected utility and communicative effort and it is related to Rate-Distortion theory (Shannon, 1948), the branch of information theory that studies the effect of limited transmission resources on communicative success (Zaslavsky et al., 2021). Its application to simple reference games has indeed demonstrated that richer decision making routines, grounded in listeners' actions and beliefs, result in human-like pragmatic behaviour (Sumers et al., 2021). Bounded rationality (Simon, 1990), which models optimal decision making under constrained cognitive resources, is a strong alternative to RSA theory but there is so far only limited evidence that it can be used to characterise language production choices (Franke et al., 2010). A third solution can be utility-based decoding algorithms—such as minimum Bayes risk (Goel and Byrne, 2000) decoding—which have been successfully used to weigh in utilities and costs during utterance selection for NLG tasks (Kumar and Byrne, 2002, 2004).

Modelling and artificially reproducing human communicative behaviour requires advanced decision making algorithms that are able to *learn from experience* efficient and effective strategies for weighing costs and utility. The learned strategies should apply both to individual utterances and to sequences of utterances: this will allow successful multi-turn planning of communicative subgoals and strategies. Reinforcement learning (RL) can naturally interact with notions of cost and utility (these can be used as learning signal for RL models, or they can be inferred by RL models from observations of human behaviour) and it has been used in combination with RSA and bounded rationality; it thus appears to be a promising avenue for the strategy learning problem.

Independent of the choice of computational model—which is an important open question—our conceptual framework can account for a variety of human behavioural patterns of communication as described in pragmatics, the field of linguistics which studies the aspects of language use that involve reasoning about context, goals, and beliefs. Let us take as an emblematic example Grice's four maxims of conversation (Grice, 1975). The maxim of *quantity*, which states that

speakers should make their contribution as informative as required for the current purposes of the exchange, can be understood as the optimisation of realisation and processing costs, C^R and C^P, while ensuring that the distance from the communicative goal is reduced. The maxim of *quality*, which is about making truthful contributions, can be thought of as the result of minimising decoding cost C^D and maximising the probability of achieving the communicative goal. The maxim of *relation*, stating that speakers should provide information that is relevant to the exchange, can be seen as a way to ensure that production and comprehension costs are always balanced by gains in positive utility. Finally, the maxim of *manner* states that speakers should avoid obscurity of expression, ambiguity, and strive for brief and orderly contributions. This can be easily understood as the optimisation of realisation and processing cost, C^R and C^P, given fixed encoding and decoding costs C^E and C^D. Grice never intended these maxims as a set of rules speakers constantly follow. When the maxims are *flouted* or *violated*, listeners can still infer communicative intents (Grice, 1975). For example, although it seems to disobey the maxim of relevance, answering 'Would you like to go to the cinema?' with 'Sorry, I am busy tonight' allows the listener to infer a negative answer to the original question. At the same time other communicative goals are achieved—e.g., the speaker is then perceived as polite.

12.6 Conclusion

This chapter presented a conceptual framework for the computational modelling of language production which relies on three central notions: communicative goals, production and comprehension costs, and joint utility. I have defined these notions formally and demonstrated their application to two realistic communication scenarios, providing examples for the modelling of goals, costs, and utility with neural models. I have further argued for the framework's ability to account for a variety of pragmatic patterns of communicative behaviour, highlighting the importance of the development of new complex decision making algorithms that learn to reproduce human-like production strategies from experience.

Chapter 13

Conclusion

The ability to use written and spoken language to transmit information is a hallmark of human intelligence. We use it to communicate facts and emotions, to coordinate joint activities, and thus to change the state of the environment in which we live and interact. A central tenet of this thesis is that human linguistic behaviour can be studied and replicated using a certain class of computational models, artificial neural networks (McClelland et al., 1986). In the last decade, deep neural networks (LeCun et al., 2015), in particular, have been used to reproduce increasingly complex aspects of human perception and behaviour, contributing to advancements in computer vision and natural language processing. Neural models that can process and produce human language are one remarkable success story: in this day and age, neural language models can automatically produce texts hardly distinguishable from those written by humans (e.g., Brown et al., 2020b), and they can engage with humans in fluent text-based language interactions (e.g., Thoppilan et al., 2022).

While neural network models of language were originally designed with the goal of studying human linguistic cognition, artificial neural architectures are mostly known and appreciated nowadays for providing a backbone for formidable AI technologies and for being commercialised as tools for a wide, general audience. This thesis has taken a different perspective. Through a series of studies on language comprehension and language production, we have investigated whether artificial neural networks—beyond being useful for search engines, chat interfaces, and content creation—can serve as accurate computational simulations of human language use, and thus as a new core methodology for the language sciences.

In each of the preceding chapters (except those dedicated to providing background), we have drawn conclusions pertaining to the respective studies and engaged in focused discussions involving promising avenues for future work. We will not delve into those here. This final chapter serves the purpose of summarising the main contributions within the three main parts of the thesis. Furthermore, it briefly addresses the scientific implications of these contributions.

Word usage. We investigated speakers' interpretation of words and the evolution of word usage over time. Our findings indicate that contemporary language models can infer contextually appropriate interpretations for diverse usages of the same word, akin to how human readers comprehend these usages. Through the integration of an artificial attention mechanism that operates across extended contextual sequences and an extensive phase of cross-situational learning using vast amounts of texts, current neural language models serve as highly generalisable engines for lexical interpretation; and they offer distinct access to the first-order and second-order co-occurrence statistics of word usage (Schütze, 1998). Within this framework, we presented two novel methods to engage with these models for obtaining word representations: the collection and analysis of the neural representations generated during the model's processing of usage examples, and the direct input of natural language instructions to induce human-readable word definitions. Both approaches hold significant relevance for examining shifts and variations in word usage across the temporal and spatial dimensions.

Utterance comprehension. Employing neural language models, we simulated the process of comprehending utterances and examined how speakers' expectations over comprehension behaviour shape the way in which information is communicated throughout texts and dialogues. This exploration served to scrutinise psycholinguistic theories concerning the rational use of the communication channel (Genzel and Charniak, 2002; Jaeger and Levy, 2007). Leveraging the capacity of contemporary neural language models to condition their probability estimates on extensive contextual sequences, we revisited the hypothesis that the pace of human information transmission remains constant, or at least uniform, throughout communication episodes, encompassing both texts and dialogues. These studies yielded fresh empirical evidence in support of information rate constancy theories in textual contexts. However, they also introduced reservations regarding the applicability of the classic information-theoretic model to naturalistic dialogue. In this domain, considerations of cost-efficiency appear to bear significant influence alongside rationality, as supported by our new findings pertaining to the facilitating effect of repeated constructions in dialogue utterances.

Utterance production. Finally, we assessed the efficacy of combining neural language models with next-word sampling algorithms, collectively referred to as 'neural generators', to emulate speakers' language production behaviour. Subsequently, these generators were used to predict elements of comprehension behaviour, such as utterance acceptability and reading times, which are known to be driven by expectations over upcoming productions. In particular, we introduced a statistical framework for the quantification of sequence-level uncertainty within these generators. Our observations demonstrated that the statistical properties of the generator output space—a window into their representation of uncertainty—

closely align with human language productions. Leveraging these novel, interpretable estimates of sequence-level uncertainty, we designed a measure of utterance predictability with substantial psychometric predictive power. Despite these encouraging results, neural language generators still lack the capacity to replicate certain human-like strategies in utterance production that we categorise as 'pragmatic', i.e., communicatively effective while also being cost-efficient. The core segment of this thesis concluded with a reflection on potential pathways for inducing the emergence of pragmatic production strategies within neural models. This reflection was complemented by a formal framework that succinctly captures the underlying perspective of this thesis, thus serving as a foundational reference for the development of more human-like artificial simulations of language usage.

An integral aspect that has been woven through the preceding chapters, and merits renewed emphasis in this concluding section, pertains to the compelling nature of neural simulations of human language use. Beyond their role in empirically scrutinising linguistic theories, neural language models can be considered as computational theories themselves. While the distinction between these stances is subtle (Baroni, 2022), the key difference lies in the fact that neural models as a theoretical construct (i) require minimal assumptions beyond the notion that the language faculty is embedded within neural connection strengths and acquired through experiential interactions, and (ii) possess the attributes of executability and quantitative verifiability. These combined attributes hold substantial promise for the language sciences, which often contend with formal, verifiable, yet undergeneralising, or conversely, informal, unprovable, and overgeneralising theories of language use.

Appendix A

Word usage

This is the appendix for Part 1. In particular, it provides supplementary information for the study presented in Chapter 5.

A.1 Preliminary analysis of usage examples

In Section 5.2.1, we present three corpora of human-written definitions and report their main statistics in Table 5.2, including mean and standard deviation of usage example length. Because the length of usage examples has been shown to affect the quality of generated definitions (Almeman and Espinosa Anke, 2022), in a preliminary analysis, we compare the length distributions of usage examples in the corpora of definitions as well as in the English DWUGs (Schlechtweg et al., 2021). Figures A.1-A.4 show the length distributions of the four datasets. We also measure the correlation between definition quality (as measured with NLG metrics: BERTScore, BLEU, NIST) and (i) the length of usage examples, (ii) the absolute position of the target word in the examples, and (iii) the target word's relative position in the examples. Tables A.1 and A.2 show the correlation coefficients.

A.2 Prompt selection

As briefly discussed in Section 5.3, in preliminary experiments, we use the pretrained Flan-T5 Base model (250M parameters; Chung et al., 2022) to select a definition generation prompt among 8 alternative verbalisations. These are a combination of four different instruction strings ('Define w', 'Define the word w', 'Give the definition of w', 'What is the definition of w?) and two ways of concatenating instructions to usage examples—i.e., either prepending them or appending them. Tables A.5-A.6 (placed at the end of this Appendix for convenience) show

	Length	Rel. Position	Abs. Position	BERT-F1	BLEU	NIST
Length	1.000	-0.123	0.575	0.067	0.076	0.045
Rel. Position	-0.122	1.000	0.626	0.053	0.075	0.062
Abs. Position	0.576	0.626	1.000	0.129	0.159	0.111
BERT-F1	0.067	0.053	**0.129**	1.000	0.121	0.095
BLEU	0.076	0.075	**0.159**	0.121	1.000	0.822
NIST	0.045	0.062	**0.111**	0.095	0.822	1.000

Table A.1: Correlations between properties of the usage examples and the quality (BERTScore, BLEU, NIST) of the definitions generated by Flan-T5 Base for WordNet. The prompt used is 'What is the definition of w?' (post). The maximum context size is set to 512.

	Length	Rel. Position	Abs. Position	BERT-F1	BLEU	NIST
Length	1.000	-0.041	0.616	0.020	0.040	0.017
Rel. Position	-0.041	1.000	0.675	0.046	0.020	0.024
Abs. Position	0.616	0.675	1.000	0.029	0.017	0.007
BERT-F1	0.020	0.046	0.029	1.000	0.283	0.277
BLEU	0.040	0.020	0.017	0.283	1.000	0.687
NIST	0.017	0.024	0.007	0.277	0.687	1.000

Table A.2: Correlations between properties of the usage examples and the quality (BERTScore, BLEU, NIST) of the definitions generated by Flan-T5 Base for Oxford. The prompt used is 'What is the definition of w?' (post). The maximum context size is set to 512.

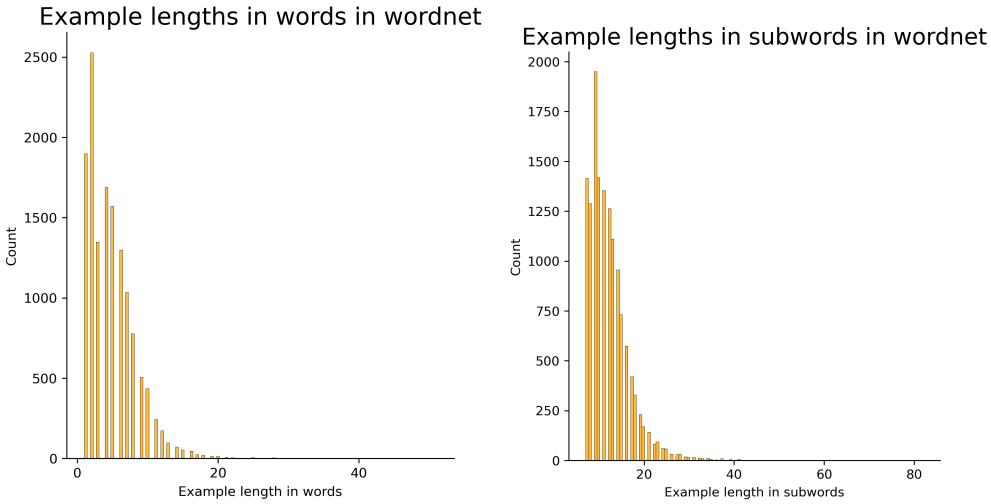


Figure A.1: Length distribution of usage examples in WordNet.

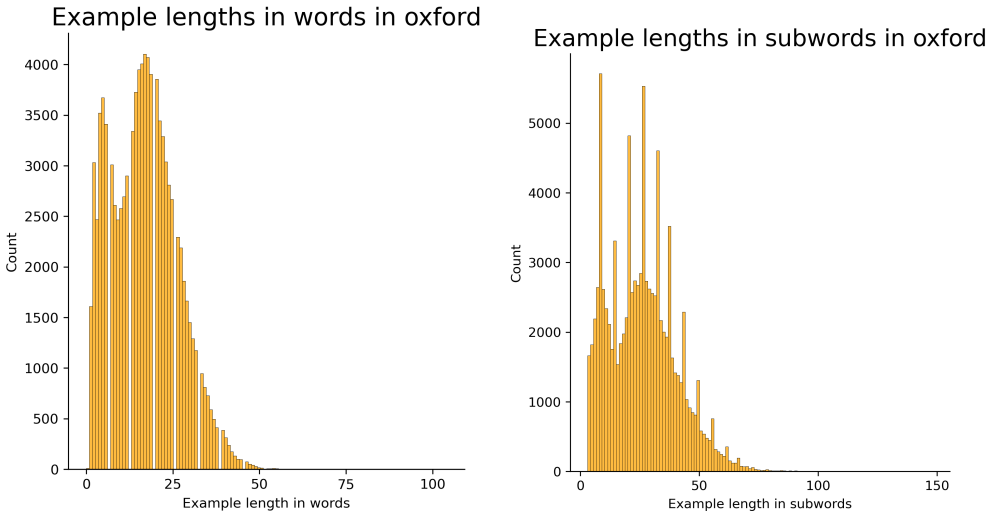


Figure A.2: Length distribution of usage examples in Oxford.

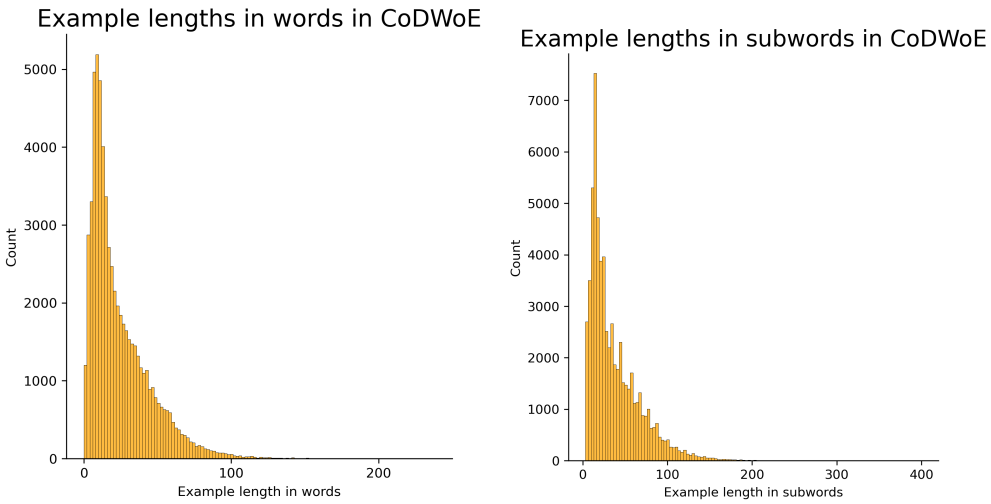


Figure A.3: Length distribution of usage examples in CoDWoE.

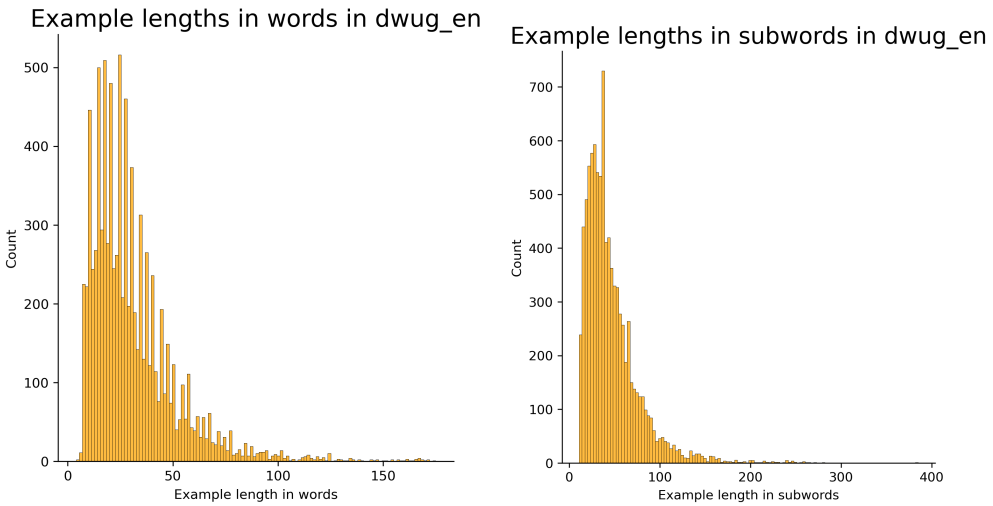


Figure A.4: Length distribution of usage examples in the English DWUGs.

the results of our experiments. In the tables, the strings 'pre' and 'post' refer to the concatenation method (prepending or appending the instruction), the numbers 128, 256, and 512 refer to the maximum length of the usage examples provided to Flan-T5 (in sub-words), and 'filter' refers to the decoding strategy of always avoiding the target word (definiendum).

A.3 Additional results: Other models and model variants

We evaluate T5 (base and XL) and Flan-T5 (base, large, and XL) under the same generalisation conditions presented for Flan T5 XL in Chapter 5 (Section 5.3.1). Results for FlanT5-XL are reported in that chapter, in Table 5.4. Here, in Table A.3, we report results for all models and model variants.

A.4 Additional examples of generated definitions and sense labels

Some definitions generated by Flan-T5 XL manage to capture very subtle aspects of the contextual lexical meaning. These, for example, are usage of 'word', accompanied by contextualised definitions:

i. 'There are people out there who have never heard of the Father, Son and Holy Spirit, let alone the **Word** of God.': *'The Bible'*

		WordNet			***Oxford***		
Model	**Test**	BLEU	ROUGE-L	BERT-F1	BLEU	ROUGE-L	BERT-F1
Huang et al. (2021)	*Unknown*	32.72	-	-	**26.52**	-	-
T5 base	Zero-shot (task shift)	2.01	8.24	82.98	1.72	7.48	78.79
T5 base	Soft domain shift	9.21	25.71	86.44	7.28	24.13	86.03
Flan-T5 base	Zero-shot (task shift)	4.08	15.32	87.00	3.71	17.25	86.44
Flan-T5 base	In-distribution	8.80	23.19	87.49	6.15	20.84	86.48
Flan-T5 base	Hard domain shift	6.89	20.53	87.16	4.32	17.00	85.88
Flan-T5 base	Soft domain shift	10.38	27.17	88.22	7.18	23.04	86.90
Flan-T5 large	Soft domain shift	14.37	33.74	88.21	10.90	30.05	87.44
T5 XL	Zero-shot (task shift)	2.05	8.28	81.90	2.28	9.73	80.37
T5 XL	Soft domain shift	**34.14**	**53.55**	91.40	18.82	38.26	88.81
Flan-T5 XL	Zero-shot (task shift)	2.70	12.72	86.72	2.88	16.20	86.52
Flan-T5 XL	In-distribution	11.49	28.96	88.90	16.61	36.27	89.40
Flan-T5 XL	Hard domain shift	29.55	48.17	91.39	8.37	25.06	87.56
Flan-T5 XL	Soft domain shift	**32.81**	**52.21**	**92.16**	18.69	**38.72**	**89.75**

Table A.3: Results of the definition generation experiments.

ii. 'Good News Bible Before the world was created, the **Word** already existed; he was with God, and he was the same as God.': *'(christianity) the second person of the trinity'*

iii. 'It was in that basement that I learned the skills necessary to succeed in the difficult thespian world-specifically, get up on stage, say my **words**, get off the stage-skills...': *'The dialogue of a play.'*

Interesting insights can be drawn from how the embeddings of the generated definitions are located in the vector space. Figure A.5 shows PCA projections of definition embeddings for usages of the words 'chef' and 'lass' from the English DWUG. Colours represent sense clusters provided in the DWUG, and the legend shows most prototypical definitions for each sense generated by our best system (singleton clusters are ignored). The large star for each sense corresponds to its sense label (as opposed to smaller stars corresponding to other definitions not chosen as the label).

For the word 'chef', there are two sense clusters, for which an identical definition is chosen (*'A commander'*). This most probably means that these clusters should in fact be merged together, or that they are in the process of splitting (see also Section 5.6 in Chapter 5). These two senses are (not surprisingly) much closer to each other than to the definitions from the *'professional cook'* sense. For the word 'lass', it is interesting how separate is a small bluish group of definitions in the bottom right corner of the plot, where the target form is actually 'lassi'. The fine-tuned Flan-T5-XL model defined this group as *'A cold drink made from milk curdled by yogurt'*, which is indeed what 'lassi' is (ignoring minor details).

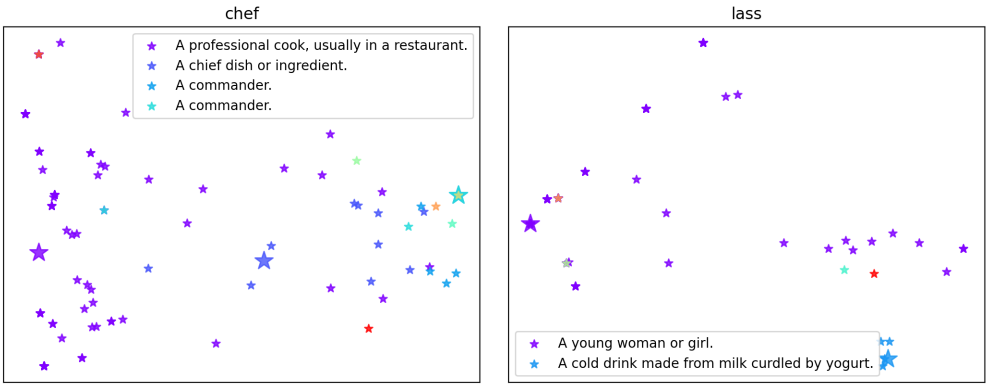


Figure A.5: PCA projections of definition embeddings for two target words from English DWUG.

A.5 Human evaluation guidelines

'You are given a spreadsheet with four columns: **Targets**, **Examples**, **System 1** and **System 2**. In every row, we have one target English word in the **Targets** column and five (or less) example usages of this word in the Examples column. Usages are simply sentences with at least one occurrence of the target word: one usage per line.

Every row is supposed to contain usages where the target word is used in the same sense: this means that for ambiguous words, there will be multiple rows, each corresponding to a particular sense. This division into senses is not always 100% correct, but for the purposes of this annotation effort, we take it for granted. Note that the five example usages in each row are sampled randomly from a larger set of usages belonging to this sense.

System 1 and System 2 are computational models which produce human-readable labels or definitions for each sense of a target word. They employ different approaches, and your task is to compare and evaluate the labels generated by these two systems. Note that in each row, the names 'System 1' and 'System 2' are randomly assigned to the actual generation systems.

The generated sense labels are supposed to be useful for historical linguists and lexicographers. Thus, they must be:

i. **Truthful**: i.e., should reflect exactly the sense in which the target word is occurring in the example usages. Ideally, the label should be general enough to encompass all the usages from the current row, but also specific enough so as not to mix with other senses (for poly-semantic target words).

ii. **Fluent**: i.e., feeling like natural English sentence or sentences, without grammar errors, utterances broken mid-word, etc

You have to fill in the **Judgements** column with one of six integer values:

- **0**: both systems are equally bad for this sense
- **1**: System 1 is better, but System 2 is also OK
- **11**: System 1 is better, and System 2 is bad
- **2**: System 2 is better, but System 1 is also OK
- **22**: System 2 is better, and System 1 is bad
- **3**: both systems are equally good for this sense

Some rows are already pre-populated with the **3** judgement, because the sense labels generated by both systems are identical. We hypothesise that this most probably means that both labels are equally good. Please still have a look at these identical labels and change **3** to **0** in case you feel that in fact they are equally bad.'

A.6 Clustering embedding spaces

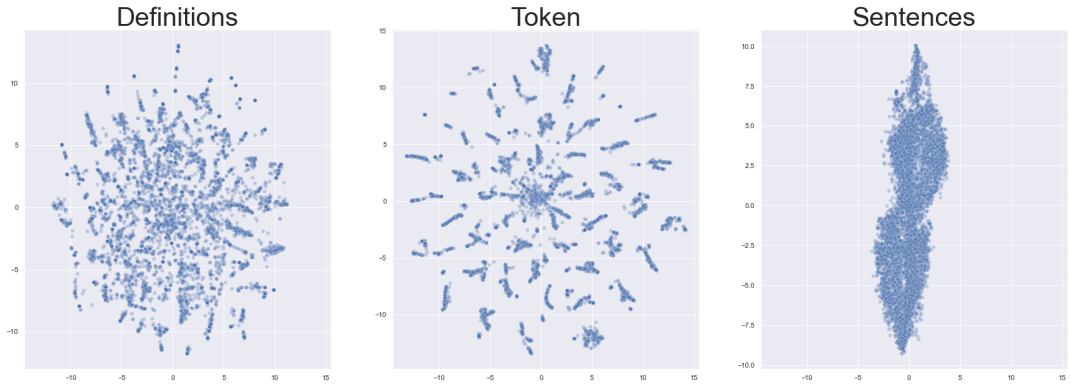


Figure A.6: T-SNE projection of each embedding space, RoBERTa-Large model.

We constructed three types of embedding spaces; (i) contextualised token embeddings, (ii) sentence embeddings, and (ii) definition embeddings. We did so for two language models: RoBERTa-large and DistilRoBERTa. Since we cluster the embedding spaces for each target word individually, we obtain different optimal number of clusters for each target word. Table 5.6 (in Chapter 5) displays the average results over all target words.

We observe that the optimal number of clusters k is substantially higher for the definition embedding spaces for both RoBERTa-large and DistilRoBERTa. However, this is an artefact of the data: since some distinct usages yield identical definitions for a target word, the definition space oftentimes consist of less

Model	**Representation**	**Sep. ↑**	**Coh. ↓**	**Ratio ↑**
RoBERTa-large	Sentence	0.017	0.013	1.248
	Token	0.042	0.034	1.272
	Definitions	0.008	0.006	1.349
DistilRoBERTa	Sentence	0.665	0.592	1.126
	Token	0.591	0.477	1.258
	Definitions	0.705	0.509	1.397

Table A.4: Separation score, cohesion score, and separation-cohesion ratio for each embedding space; average over all target words from the English DWUGs.

distinct data points, which greatly impacts the average silhouette scores. Future work should point out what clustering methods are most applicable to definition embedding spaces. Still, this decrease in data points confirms how the definition embedding space could represent usages at a higher level of abstraction, collapsing distinct usages into identical representations.

Figure A.6 displays the T-SNE projections of each of the three embedding spaces of RoBERTA-large. As for Distil-RoBERTa, the definition embedding space appears to have spacial properties that are more similar to contextualised *token* embedding spaces than to *sentence* embedding spaces: the definition embeddings are more separated than the sentence embeddings, and are cluttered in a similar manner as the token embeddings.

Table A.4 shows the average inter- and intra-cluster dispersion values of the clusters as labelled by the English DWUGs (Schlechtweg et al., 2021). These are calculated for the token, sentence and definition embeddings of both RoBERTa-large and Distil-RoBERTa.

Configuration	**BLEU**	**NIST**	**BERT-F1**
what is the definition of <trg>? post 256	0.0985	0.1281	0.8700
what is the definition of <trg>? post 512	0.0985	0.1281	0.8700
give the definition of <trg> post filter	0.0719	0.1520	0.8560
give the definition of <trg> post 256	0.0629	0.1563	0.8522
give the definition of <trg> post 512	0.0629	0.1563	0.8522
define the word <trg> post 512	0.0462	0.0972	0.8512
define the word <trg> post 256	0.0462	0.0972	0.8512
give the definition of <trg>: pre 256	0.0446	0.1123	0.8495
what is the definition of <trg>? pre 512	0.0403	0.0705	0.8495
give the definition of <trg>: pre 512	0.0446	0.1123	0.8495
what is the definition of <trg>? pre 256	0.0403	0.0703	0.8494
define the word <trg>: pre 512	0.0313	0.0615	0.8481
define the word <trg>: pre 256	0.0313	0.0618	0.8480
define <trg> post 512	0.0275	0.0583	0.8475
define <trg> post 256	0.0275	0.0583	0.8475
define <trg>: pre 512	0.0195	0.0411	0.8453
define <trg>: pre 256	0.0195	0.0409	0.8453

Table A.5: Prompt selection results on WordNet.

Configuration	**BLEU**	**NIST**	**BERT-F1**
give the definition of <trg>: pre 64	0.0680	0.1513	0.8461
what is the definition of <trg>? post 64	0.1068	0.1464	0.8458
give the definition of <trg> post 64	0.0654	0.1602	0.8374

Table A.6: Prompt selection results on CoDWoE Trial.

Configuration	**BLEU**	**NIST**	**BERT-F1**
what is the definition of <trg>? post 512	0.1232	0.1488	0.8648
what is the definition of <trg>? post 128	0.1232	0.1488	0.8648
what is the definition of <trg>? post 256	0.1232	0.1488	0.8648
what is the definition of <trg>? post oxford filter 128	0.1219	0.1398	0.8644
give the definition of <trg> post 128	0.0823	0.1793	0.8531
give the definition of <trg> post 256	0.0823	0.1793	0.8531
give the definition of <trg> post 512	0.0823	0.1793	0.8531
give the definition of <trg> post oxford filter 128	0.0763	0.1415	0.8526
what is the definition of <trg>? pre 256	0.0801	0.0966	0.8501
what is the definition of <trg>? pre 512	0.0801	0.0966	0.8501
what is the definition of <trg>? pre 128	0.0801	0.0966	0.8501
give the definition of <trg>: pre 128	0.0695	0.1313	0.8493
give the definition of <trg>: pre 256	0.0695	0.1313	0.8493
give the definition of <trg>: pre 512	0.0695	0.1313	0.8492
define the word <trg> post 128	0.0614	0.1112	0.8442
define the word <trg> post 512	0.0614	0.1112	0.8442
define the word <trg> post 256	0.0614	0.1112	0.8442
define the word <trg>: pre 256	0.0408	0.0602	0.8352
define the word <trg>: pre 512	0.0408	0.0602	0.8352
define the word <trg>: pre 128	0.0408	0.0602	0.8352
define <trg> post 256	0.0279	0.0581	0.8319
define <trg> post 128	0.0279	0.0581	0.8319
define <trg> post 512	0.0279	0.0581	0.8319
define <trg>: pre 512	0.0161	0.0237	0.8305
define <trg>: pre 256	0.0160	0.0237	0.8305
define <trg>: pre 128	0.0160	0.0237	0.8305

Table A.7: Prompt selection results on Oxford.

Configuration	**BLEU**	**NIST**	**BERT-F1**
what is the definition of <trg>? post 128	0.1138	0.2137	0.8702
give the definition of <trg> post 128	0.0826	0.2389	0.8615
what is the definition of <trg>? post 64	0.1033	0.1990	0.8595
give the definition of <trg> post 64	0.0785	0.2194	0.8520

Table A.8: Prompt selection results on CoDWoE Complete.

Appendix B

Utterance comprehension

This is the appendix for Part 2. It provides supplementary information for the methods and experiments presented in Chapters 6, 7, and 8.

B.1 Corpus excerpts

Tables B.1, B.2, and B.3 show excerpts of a Penn Treebank article, a PhotoBook dialogue, and a Spoken BNC dialogue. The article (Table B.1) is annotated with sentence positions and surprisal estimates. The dialogues (Tables B.2 and B.3) are annotated with utterance positions, speaker identifiers, and surprisal estimates.

B.2 Language models

We experiment with GPT-2 (Radford et al., 2019), an autoregressive Transformer-based (Vaswani et al., 2017) language model, and we rely on HuggingFace's implementation with default tokenizers and default parameters (Wolf et al., 2020).[1] We use the model's maximum sequence length, 1024. As the pre-trained model yields relatively high perplexity on the target corpora, we fine-tune it on 70% of each target corpus and leave out 30% of the dataset to compute the model's evaluation perplexity and to conduct our statistical analysis.[2] The training and held-out portions of the corpora are specified in the main chapters of Part 2. GPT-2 is fine-tuned for 20 epochs with a learning rate of $1e-04$ and batches of size 8. Because 20 epochs do not yield a substantial perplexity reduction for the Spoken BNC dialogues, we fine-tuned the model for 20 additional epochs. The perplexity of the pre-trained and fine-tuned models on the target corpora is reported in Table 6.3 (Chapter 6).

[1] The pre-trained model is available at `https://huggingface.co/gpt2`.

[2] We use HuggingFace's fine-tuning script `https://github.com/huggingface/transformers/blob/master/examples/pytorch/language-modeling/run_clm.py`.

Pos.	Sentence	$H(S)$	$H(S\|C)$
1	Storage Technology Corp. said it signed a letter of intent to acquire M4 Data Inc. of Britain.	3.89	3.89
2	Terms weren't disclosed.	2.26	2.11
3	Storage Technology said M4's magnetic tape storage equipment will complement its tape cartridge products.	7.64	6.55
4	M4 sells to the original equipment manufacturer market world-wide and has about $20 million in annual sales.	5.75	5.50
5	A Storage Technology spokesman said the transaction should be completed in one to two months.	4.45	3.81

Table B.1: A Penn Treebank article (document id: 15) annotated with utterance positions (Pos.) and surprisal estimates.

For our surprisal estimates, we include a special sentence beginning symbol as a basic contextual cue, but its surprisal is not computed. Furthermore, for the dialogue corpora, we try prepending input utterances with dialogue turn cues ('A: ', 'B: ') as a hint to the language models that the data is conversational. This modification of the input text does not consistently reduce the models' perplexity scores.

B.2.1 Transformer-XL

Although excluding high sentence positions is in line with prior work measuring decontextualised surprisal (e.g., Genzel and Charniak, 2002, 2003; Xu and Reitter, 2018), we have tried to substitute GPT-2 with the Transformer-XL language model (Dai et al., 2019) because of the latter's unlimited context window size. In spite of its larger window, however, Transformer-XL yields higher perplexity than GPT-2 on all corpora. Moreover, to make fine-tuning computationally feasible, we had to limit the context window size to values close to 1024; this is likely to make the model unable to use very long-distance dependencies at inference time, making it more similar but less performant than GPT-2. Indeed, Transformer-XL models fine-tuned with a fixed context size of 1024 yield higher perplexity than the corresponding fine-tuned GPT-2 models.

Pos.	**Id.**	**Utterance**	$H(S)$	$H(S\|C)$
1	A	Do you have a boy in an orange shirt jumping near a boat ?	3.64	3.64
2	B	Yes.	4.86	5.12
3	A	do you have a miltary boat that shows a man climbing a ladder?	4.25	4.03
4	B	I don't have that one.	1.28	1.47
5	B	I have a woman in a white hat, red boat and blue life vest.	3.62	3.29
6	A	I dont have that	2.69	2.87
7	A	do you have a man in a vest and tie at night against the railing	4.64	4.30
8	B	Yes.	4.86	5.20
9	A	any other questions?	4.05	3.84
10	A	do you see two ladies with a panda bear doll on a boat ?	4.87	4.82
11	B	Yes.	4.86	3.85
12	A	do you see the military man climbing the ladder from the raft in a helmet	4.85	4.42
13	B	Yep. I have that one, too.	2.77	2.32
14	A	do you see a lady in kayak and whit hat red kayak?	4.31	3.97
15	B	I don't have that one this time.	1.51	1.33
16	A	do you have questions?	4.14	4.52
17	B	I have an Asian sitting near several stacks of wood.	6.08	5.63
18	A	no i dont have that	2.78	2.70

Table B.2: The first two rounds of a PhotoBook dialogue (dialogue id: 1861), annotated with utterance positions (Pos.), speaker identifier (Id.), and surprisal estimates.

Pos.	**Id.**	**Utterance**	$H(S)$	$H(S\|C)$
1	S0018	so how come you're back so early? I thought you had a tennis lesson	3.50	3.50
2	S0019	oh well so did I	5.75	5.74
3	S0019	and having made the arrangement with last Tuesday carefully explaining to him that I couldn't do tomorrow because of the funeral he said well okay I can do twelve o'clock on Monday fine so I toddles along at twelve o'clock today to be told that 's on a course at	3.64	3.79
4	S0018	oh no	5.28	5.22
5	S0019	but had obviously not bothered to write it down	6.08	5.83
6	S0018	so he'd just completely forgotten you?	5.72	5.13
7	S0019	yes in a word	7.36	6.48
8	S0018	Did you phone him?	6.05	6.36
9	S0019	no I didn't I allowed myself a little bit of time to not be quite so cross and I had er half an hour with well more than half an hour three-quarters of an hour with one of the other coaches there	3.95	3.71
10	S0018	what he just happened to be free?	5.71	6.06

Table B.3: The first ten turns of a Spoken BNC dialogue (dialogue id: SVNL), annotated with utterance positions (Pos.), speaker identifier (Id.), and surprisal estimates.

B.2.2 Effects of fine-tuning on the dialogue corpora

Here, we give an intuition of the main effects of fine-tuning GPT-2 on dialogue corpora, focusing on PhotoBook and Map Task.

The following are the main effects of fine-tuning GPT-2 on PhotoBook dialogues:

- The fine-tuned model is less surprised by utterance types that are frequent in the corpus; the least surprising expressions are *I have*, *I don't have that one*, *I don't have that*, *No, I don't have that one*. For the pre-trained model, on the other hand, the least surprising expressions are more generic: *No, I don't think so*, *I'm not sure*, *I don't*, *What do you think?*.
- Among the most surprising utterances for the pre-trained model are some that are specific to PhotoBook games: *submit bye*, *loading may be frozen*. For these two utterances, e.g., surprisal decreases by 1/4 and 1/3 respectively after fine-tuning.
- Written chat language becomes less surprising: e.g., the surprisal for *kk done* decreases by one third.
- Utterances at first dialogue positions become in general less surprising. The decrease in surprisal for greetings is not always consistent: e.g., the surprisal for *hi* and *hey there* decrease by one third and one seventh respectively.

These are the main effects of fine-tuning GPT-2 on MapTask dialogues:

- While the pre-trained model assigns high surprisal to utterances that contain disfluencies, this is not the case for the fine-tuned model.
- Backchannels also become less surprising with fine-tuning: the surprisal of, e.g., *okay*, *mmhmm*, *well*, *right*, *erm*, *yeah*, *no*, *aye* decreases by 25% to 75%.
- With fine-tuning, GPT-2 doesn't only get used to features of transcribed speech: expressions that refer to MapTask landmarks also become more likely (e.g., *the rapids*, *a rope bridge*, *the gold mine*).
- Simple spatial indications (*towards the bottom left-hand corner*, *on the left-hand side*) are among the utterances with the lowest surprisal.

B.3 Replication study: Surprisal constancy in newspaper articles

We use the Wall Street Journal part of the Penn Treebank, divided into a training set (section 0–20) and a test set (sections 21–24). The training set contains

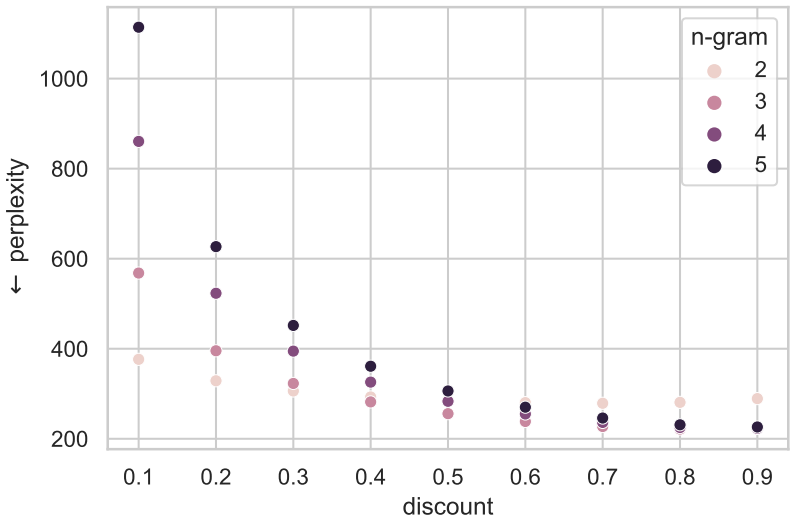


Figure B.1: Perplexity on Penn Treebank test set obtained by n-gram language models with Kneser-Ney smoothing and interpolation.

41,128 sentences (Keller (2004) reports 42,075 sentences), the test set 8,594 (Keller (2004) reports 7,133). Each article is treated as a separate text, and sentence positions are computed by counting from the beginning of the article. The sentence positions in the test set vary between 1 and 118 (Keller (2004) reports 1-149). The n-gram probabilities are computed by Keller (2004) using a language model with smoothing by absolute discounting, whereas Genzel and Charniak (2002) do not report the specifics of their language model. We rely on NLTK's implementation of an n-gram language model with interpolated Kneser-Ney smoothing (Bird et al., 2009). We train n-gram language models with $n \in (2, 3, 4, 5)$ and with discount values $n \in (0.1, 0.2, \ldots, 0.9)$ on the training set and select the language model with the lowest perplexity on the test set. The best language model is the 3-gram model with a discount value of 0.8, which achieves a perplexity of 335.80 on the test set. The perplexity obtained using NLTK's evaluation script is 221.57 (Figure B.1) as it is calculated by taking into account beginning and end of sentence symbols.

We use the n-gram language model as well as GPT-2 to estimate the surprisal of all sentences in the test set and measure the correlation with sentence position. In Genzel and Charniak's original work (2002), the correlation between sentence position and sentence surprisal is computed by binning the sentence surprisal data points based on their sentence position. Correlation is measured between sentence position indices 1-25 and the average sentence surprisal estimated for the respective sentence position. Keller (2004) also measures the raw correlation between all sentence position-surprisal pairs, without binning. Neither work re-

	Cut-off = 25	**Cut-off = 76**	**Cut-off = ∞**
Raw data	τ	τ	τ
3-gram (Keller, 2004)	0.060**	0.081**	0.071**
3-gram (ours)	0.076**	0.081**	0.079**
GPT-2 pre-trained	0.032**	0.055**	0.054**
GPT-2 fine-tuned	0.070**	0.080**	0.080**
Binned data	τ	τ	τ
3-gram (Keller, 2004)	0.639**	0.243**	0.135
3-gram (ours)	0.733**	0.109	0.118
GPT-2 pre-trained	0.533**	0.512**	0.077
GPT-2 fine-tuned	0.693**	0.387**	0.119

Table B.4: Kendall's rank correlation between sentence surprisal and sentence position for the Penn Treebank test set. Significance: '**' $p < 0.001$, '*' $p < 0.01$, ' ' $p \geq 0.05$.

	Cut-off = 25	**Cut-off = 76**	**Cut-off = ∞**
Raw data	τ	τ	τ
3-gram (Keller, 2004)	0.078**	0.093**	0.081**
3-gram (ours)	0.082**	0.087**	0.087**
GPT-2 pre-trained	0.034**	0.054**	0.054**
GPT-2 fine-tuned	0.077**	0.084**	0.084**
Binned data	τ	τ	τ
3-gram (Keller, 2004)	0.671**	0.147	0.170**
3-gram (ours)	0.740**	0.099	0.097
GPT-2 pre-trained	0.453*	0.448**	0.101
GPT-2 fine-tuned	0.680**	0.347**	0.104

Table B.5: Kendall's rank correlation between sentence surprisal and sentence position, with sentence length partialled out, for the Penn Treebank test set. Significance: '**' $p < 0.001$, '*' $p < 0.01$, ' ' $p \geq 0.05$.

ports the correlation measure used. We use Kendall's rank correlation as it is less sensitive than Spearman's rank correlation to the large amount of ties (position-surprisal pairs with the same position index) in our data. Moreover, whereas Genzel and Charniak (2002) select a single sentence position cut-off ($c = 25$), in Keller's (2004) study three variants of the cut-off are used ($c = 25$, $c = 76$, and no cut-off). We also compute correlation at these three levels. Finally, following Keller (2004), we compute the partial correlation between sentence position and sentence surprisal, excluding the effect of sentence length. The results are reported in Tables B.4 and B.5.

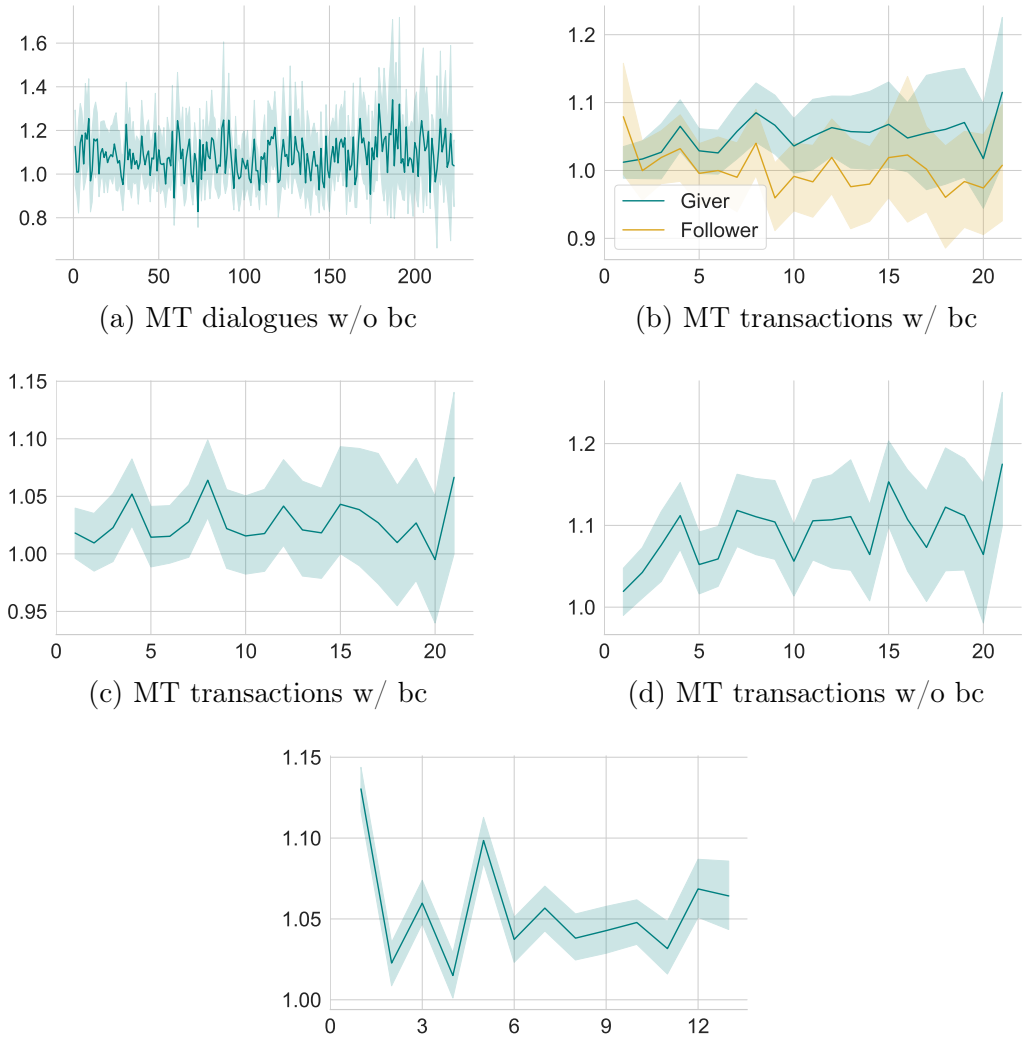
B.4 Experimental results: Utterance surprisal as a function of discourse context

Tables B.7, B.8 and B.9 summarise the results of our statistical analysis, as introduced in Chapter 7. For convenience, they are shown at the end of Appendix B. Figure B.2 (on the next page) shows the patterns of surprisal against turn position for the contextual units whose patterns are not displayed in Section 7.3.2.

Our linear mixed effect models include the logarithm of the information theoretic estimate of interest (decontextualised surprisal $H(S)$, contextualised surprisal $H(S|C)$, or context informativeness $I(S;C)$) as the response variable; the logarithm of utterance position and the logarithm of utterance length as predictors; a random intercept grouped by distinct documents/dialogues; and a document-specific random slope for utterance position and utterance length. The Random effects columns show the standard deviation of the random effects (Coeff.) and the residual standard deviation.

B.5 Extraction of repeated constructions

We define a limited specific vocabulary of generic nouns that should not be considered referential. The vocabulary includes: *bit*, *bunch*, *day*, *days*, *fact*, *god*, *idea*, *ideas*, *kind*, *kinds*, *loads*, *lot*, *lots*, *middle*, *ones*, *part*, *problem*, *problems*, *reason*, *reasons*, *rest*, *side*, *sort*, *sorts*, *stuff*, *thanks*, *thing*, *things*, *time*, *times*, *way*, *ways*, *week*, *weeks*, *year*, *years*. We also find all the filled pauses and exclude word sequences that consist for more than 50% of filled pauses. Filled pauses in the Spoken BNC are transcribed as: *huh*, *uh*, *erm*, *hm*, *mm*, *er*.



(a) MT dialogues w/o bc

(b) MT transactions w/ bc

(c) MT transactions w/ bc

(d) MT transactions w/o bc

(e) PB rounds

Figure B.2: surprisal (y axis) against turn position (x axis) in MapTask (MT) dialogues and transactions—with or without backchannels (w/ bc and w/o bc, respectively)—and PhotoBook (PB) dialogue rounds. Position is cut off at mean + 1 sd. Bootstrapped 95% confidence bands.

B.6 Adaptive language model

B.6.1 Fine-tuning

We fine-tune the 'small' variants of GPT-2 (Radford et al., 2019) and DialoGPT (Zhang et al., 2020b) on our fine-tuning split of the Spoken BNC (see Section 8.3) using HuggingFace's implementation of the models with default tokenizers and parameters (Wolf et al., 2020). Dialogue turns are simply concatenated; we experimented with labelling the dialogue turns (i.e., *'A: utterance 1, B: utterance 2'* and found that this leads to higher perplexity. The fine-tuning results for both models are presented in Table B.6. We fine-tune the models and measure their perplexity using Huggingface's fine-tuning script. We use early stopping over 5 epochs.[3] Sequence length and batch size vary together because they together determine the amount of memory required; more expensive combinations (e.g., 256 tokens with batch size 16) require an exceedingly high amount of GPU memory. Reducing the maximum sequence length has limited impact: 99.90% of dialogue turns have at most 128 words.

DialoGPT starts from extremely high perplexity values but catches up quickly with fine-tuning. GPT-2 starts from much lower perplexity values and reaches virtually the same perplexity as DialoGPT after fine-tuning. For the pre-trained DialoGPT, perplexity is extremely high, and the perplexity trend against maximum sequence length is surprisingly upward. These two behaviours indicate that the pre-trained DialoGPT is less accustomed than GPT-2 to the characteristics of our dialogue data. DialoGPT is trained on written online group conversations, while we use a corpus of transcribed spoken conversations between two speakers. In contrast, GPT-2 has been exposed to the genre of fiction, which contains scripted dialogues, and thus to a sufficiently similar language use. We select GPT-2 fine-tuned with a maximum sequence length of 128 and 512 as our best two models; these two models (which we now refer to as *frozen*) are used for the adaptive learning rate selection procedure (Section B.6.2).

B.6.2 Learning rate selection

To find the appropriate learning rate for on-the-fly adaptation (see Section 8.4.2), we randomly select 18 dialogues D from the analysis split of the Spoken BNC and run an 18-fold cross-validation for a set of six candidate learning rates: 1×10^{-5}, 1×10^{-4}, ..., 1. We fine-tune the model on each dialogue using one of these

[3]The number of epochs (5) has been selected in preliminary experiments together with the learning rate (1×10^{-4}). In these experiments—which we ran for 40 epochs—we noticed that the 1×10^{-4} learning rate offers the best tradeoff of training time and perplexity out of four possible values: 1×10^{-2}, 1×10^{-3}, 1×10^{-4}, 1×10^{-5}. We obtained insignificantly lower perplexity values with a learning rate of 1×10^{-5}, with significantly longer training time: 20 epochs for GPT-2 and 28 epochs for DialoGPT.

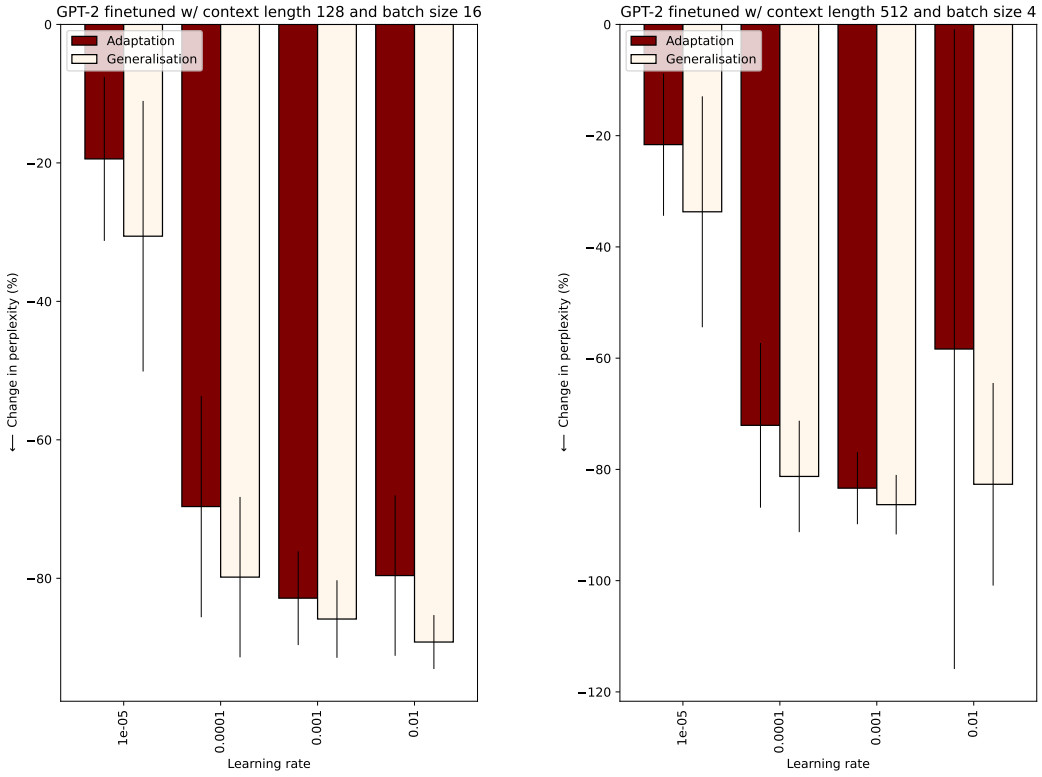
Model	Learning rate	Max seq. length	Batch size	Best epoch	Ppl fine-tuned	Ppl pre-trained
DialoGPT	0.0001	128	16	3	23.21	7,091.38
DialoGPT	0.0001	256	8	4	22.26	12,886.92
DialoGPT	0.0001	512	4	4	21.73	21,408.32
GPT-2	0.0001	128	16	4	23.32	173.76
GPT-2	0.0001	256	8	3	22.21	159.23
GPT-2	0.0001	512	4	3	21.55	149.82

Table B.6: Fine-tuning results for GPT-2 and DialoGPT on our fine-tuning split of the Spoken BNC.

learning rate values, and compute perplexity change 1) on the dialogue itself (to measure *adaptation*) as well as 2) on the remaining 17 dialogues (to measure *generalisation*). We set the Transformer's context window to 50 to reproduce the experimental conditions presented in Section 8.4.1.

More precisely, for each dialogue $d \in D$, we calculate the perplexity of our two frozen models (Section B.6.1) on d and $D \setminus \{d\}$ (which we refer to as $ppl_{before}(d)$ and $ppl_{before}(D)$, respectively). Then, we fine-tune the models on d using the six candidate learning rates, and measure again the perplexity over d and $D \setminus \{d\}$ (respectively, $ppl_{after}(d)$ and $ppl_{after}(D)$). The change in performance is evaluated according to two metrics: $\frac{ppl_{after}(d) - ppl_{before}(d)}{ppl_{before}(d)}$ measures the degree to which the model successfully adapts to the target dialogue; $\frac{ppl_{after}(D) - ppl_{before}(D)}{ppl_{before}(D)}$ measures whether fine-tuning on the target dialogue causes any loss of generalisation.

The learning rate selection results are presented in Figure B.3. We select 1×10^{-3} as the best learning rate and pick the model fine-tuned with a maximum sequence length of 512 as our best model. The difference in perplexity reduction (both adaptation and generalisation) is minimal with respect to the model fine-tuned with a maximum sequence length of 128, but since the analysis split of the Spoken BNC contains turns longer than 128 tokens, we select the 512 version. Similarly to van Schijndel and Linzen (2018), we find that fine-tuning on a dialogue does not cause a loss in generalisation but instead helps the model generalise to other dialogues. Unlike van Schijndel and Linzen (2018), who used LSTM language models, we find that learning rates larger than 1×10^{-1} cause back-propagation to overshoot, even within a single dialogue. In Figure B.3, the bars for 1×10^{-1} and 1 are not plotted because the corresponding data contains infinite perplexity values (due to numerical overflow). The selected learning rate, 1×10^{-3}, is a relatively low learning rate for on-the-fly adaptation but it is still higher than the best learning rate for the entire dataset by a factor of 10.



(a) (b)

Figure B.3: The adaptation and generalisation performance (defined in Section B.6.2) with varying learning rate.

Listing B.1: Linear mixed effect model for Facilitating Effect

```
MODEL INFO:
Observations: 46399
Dependent Variable: Facilitating Effect
Type: Mixed effects linear regression

MODEL FIT:
AIC = 99197.283, BIC = 99302.224
Pseudo-R^2 (fixed effects) = 0.084
Pseudo-R^2 (total) = 0.111

FIXED EFFECTS:
```

	Est.	2.5%	97.5%	t val.	d.f.	p
(Intercept)	0.704	0.683	0.725	65.527	185.698	0.000
log Utterance Position	0.046	0.026	0.066	4.556	9274.269	0.000
log Construction Length	0.098	0.084	0.111	14.396	46372.022	0.000
log Repetition Index	0.079	0.063	0.094	10.096	45082.205	0.000
log Distance	−0.311	−0.328	−0.293	−34.571	46269.156	0.000
Previous Same Utterance	−0.099	−0.184	−0.013	−2.262	46063.723	0.024
log Rep. Index in Utterance	0.178	0.130	0.226	7.243	45765.367	0.000
PMI	−0.139	−0.154	−0.124	−18.225	45172.205	0.000
Referential	0.124	0.099	0.149	9.887	46214.616	0.000

```
p values calculated using Satterthwaite d.f.

RANDOM EFFECTS:
```

Group	Parameter	Std. Dev.
Speaker:‘Dialogue ID	(Intercept)	0.082
Dialogue ID	(Intercept)	0.090
Residual		0.701

```
Grouping variables:
```

Group	# groups	ICC
Speaker:‘Dialogue ID	368	0.013
Dialogue ID	185	0.016

```
Continuous predictors are mean-centered and scaled by 2 s.d.
```

B.7 Experimental results: The facilitating effect of construction repetition

This section of the appendix presents the main statistical model used in Chapter 8. As explained in Section 8.6.1, we fit a linear mixed effect model using facilitating effect as the response variable and including multilevel random effects grouped by dialogues and individual speakers.[4]. The fixed effects of the model, resulting from a backward stepwise selection procedure, are presented in Section 8.6.1. Non-binary predictors are log-transformed, mean-centered, and scaled by 2 sd. The final model is summarised in Listing B.1 and its coefficients are visualised in Figure B.4. We use the `lme4` and `lmerTest` R packages for this analysis.

[4]We also try grouping observations only by dialogue and only by individual speakers. The amount of variance explained (but unaccounted for by the fixed effects) decreases, so we keep the two-level random effects.

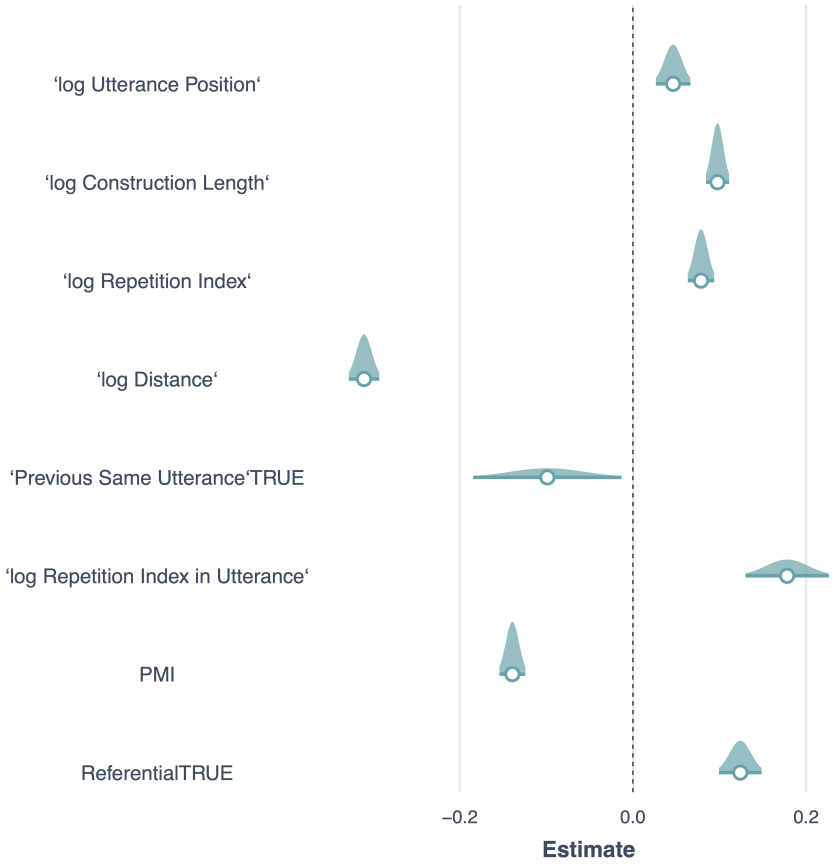


Figure B.4: Significant predictors of facilitating effect. Mixed effects linear regression, continuous predictors are mean-centred and scaled by 2 standard deviations.

		Fixed effects			Random effects	
		Estimate	**Std. Error**	**Pr(>|t|)**	**Coeff.**	**Residual**
PTB: $H(S)$	Intercept	1.966	0.025	<0.001	0.332	0.186
	Position	0.029	0.004	<0.001	0.043	
	Length	-0.125	0.006	<0.001	0.076	
PTB: $H(S\|C)$	Intercept	1.878	0.026	<0.001	0.320	0.204
	Position	0.002	0.004	0.545	0.037	
	Length	-0.107	0.007	<0.001	0.076	
PTB: $I(S;C)$	Intercept	0.711	0.048	<0.001	0.587	0.397
	Position	0.121	0.007	<0.001	0.058	
	Length	-0.173	0.013	<0.001	0.164	
PB: $H(S)$	Intercept	1.786	0.010	<0.001	0.183	0.337
	Position	0.041	0.002	<0.001	0.042	
	Length	-0.181	0.003	<0.001	0.056	
PB: $H(S\|C)$	Intercept	1.986	0.010	<0.001	0.190	0.397
	Position	-0.016	0.003	<0.001	0.039	
	Length	-0.250	0.003	<0.001	0.065	
PB: $I(S;C)$	Intercept	-1.089	0.027	<0.001	0.559	0.846
	Position	0.279	0.007	<0.001	0.134	
	Length	0.355	0.009	<0.001	0.199	
BNC: $H(S)$	Intercept	1.813	0.015	<0.001	0.144	0.287
	Position	-0.001	0.003	0.875	0.027	
	Length	-0.080	0.004	<0.001	0.038	
BNC: $H(S\|C)$	Intercept	1.729	0.025	<0.001	0.241	0.492
	Position	-0.029	0.006	<0.001	0.060	
	Length	-0.051	0.006	<0.001	0.065	
BNC: $I(S;C)$	Intercept	0.446	0.049	<0.001	0.351	1.154
	Position	0.063	0.012	<0.001	0.075	
	Length	-0.104	0.011	<0.001	0.087	

Table B.7: Results of linear mixed effect models on the Penn Treebank newspaper articles (PTB), the PhotoBook written task-oriented dialogues (PB), and the Spoken BNC open-domain dialogues (BNC).

		Fixed effects			Random effects	
		Estimate	**Std. Error**	**Pr(>|t|)**	**Coeff.**	**Residual**
		Estimate	**Std. Error**	**Pr(>|t|)**	**Coeff.**	**Residual**
MT dial. w/ bc	Intercept	0.07×10^{-2}	2.40×10^{-2}	0.98	11.44×10^{-2}	30.05×10^{-2}
	Position	-0.70×10^{-2}	0.37×10^{-2}	0.06	0.80×10^{-2}	
MT dial. w/o bc	Intercept	3.37×10^{-2}	3.20×10^{-2}	0.30	15.05×10^{-2}	26.83×10^{-2}
	Position	0.03×10^{-2}	0.57×10^{-2}	0.96	1.98×10^{-2}	
MT trans. w/ bc	Intercept	-2.95×10^{-2}	1.50×10^{-2}	0.06	7.90×10^{-2}	30.07×10^{-2}
	Position	0.03×10^{-2}	0.37×10^{-2}	0.93	0.34×10^{-2}	
	Intercept (givers)	-2.20×10^{-2}	1.49×10^{-2}	0.15	7.38×10^{-2}	28.40×10^{-2}
	Position (givers)	0.92×10^{-2}	0.46×10^{-2}	0.06	0.90×10^{-2}	
	Intercept (followers)	-5.01×10^{-2}	2.30×10^{-2}	0.04	10.88×10^{-2}	31.40×10^{-2}
	Position (followers)	0.60×10^{-2}	0.71×10^{-2}	0.41	1.50×10^{-2}	
MT trans. w/o bc	Intercept	-0.93×10^{-2}	1.61×10^{-2}	0.57	7.98×10^{-2}	26.80×10^{-2}
	Position	2.38×10^{-2}	0.49×10^{-2}	< 0.01	0.96×10^{-2}	
	Intercept (givers)	-3.78×10^{-2}	1.70×10^{-2}	0.03	8.20×10^{-2}	25.47×10^{-2}
	Position (givers)	3.46×10^{-2}	0.53×10^{-2}	< 0.01	0.19×10^{-2}	
	Intercept (followers)	9.04×10^{-2}	3.10×10^{-2}	< 0.01	13.10×10^{-2}	28.27×10^{-2}
	Position (followers)	-1.30×10^{-2}	1.38×10^{-2}	0.36	5.50×10^{-2}	

Table B.8: Results of linear mixed effect models on the MapTask data, with surprisal estimates obtained within dialogues (dial.) and transactions (trans.), with and without backchannels (bc).

		Fixed effects			Random effects	
		Estimate	**Std. Error**	**Pr(>|t|)**	**Coeff.**	**Residual**
PB dialogues	Intercept	-12.21×10^{-2}	0.90×10^{-2}	< 0.01	17.52×10^{-2}	37.66×10^{-2}
	Position	3.13×10^{-2}	0.24×10^{-2}	<0.01	3.77×10^{-2}	
PB rounds	Intercept	-0.99×10^{-2}	0.70×10^{-2}	0.16	15.14×10^{-2}	37.82×10^{-2}
	Position	-0.73×10^{-2}	0.26×10^{-2}	<0.01	3.46×10^{-2}	
PB chains	Intercept	-5.92×10^{-2}	0.36×10^{-2}	<0.01	14.86×10^{-2}	28.95×10^{-2}
	Position	1.27×10^{-2}	0.27×10^{-2}	<0.01	4.70×10^{-2}	

Table B.9: Results of linear mixed effect models on the PhotoBook data.

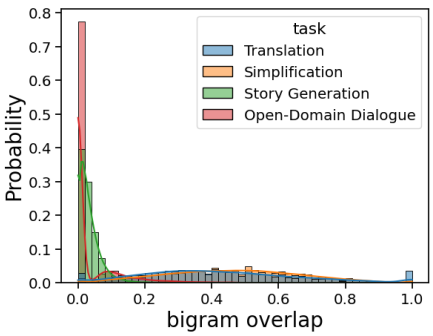
Appendix C

Utterance production

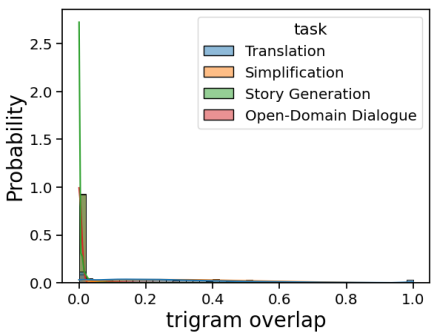
This is the appendix for Part 3. It provides supplementary information for the methods and experiments presented in Chapters 10 and 11.

C.1 Further figures on production variability

Figures C.1 and C.2 show human production variability over lexical and syntactic unigrams, bigrams, and trigrams (complementing Figure 10.2 in Chapter 10). Figures C.3 to C.5 show mean divergences across tasks, probes, and decoding algorithms (complementing Figure 10.6 in Chapter 10).

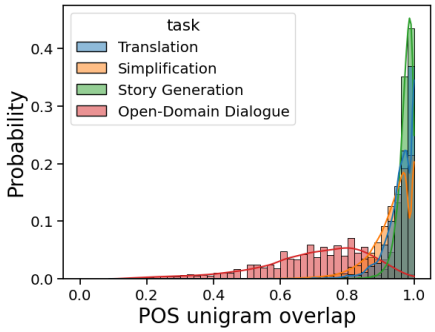


(a) Lexical variability

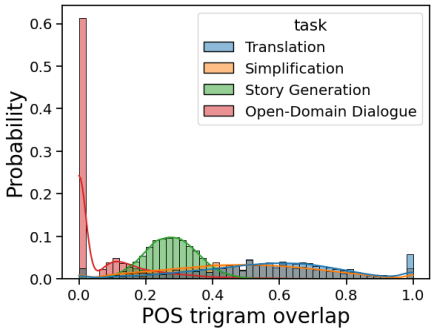


(b) Lexical variability

Figure C.1: Human production variability across four NLG tasks. The values on the x-axis are single samples of lexical, syntactic, or semantic similarity between two productions for each input (see Section 10.2). Probability mass on the right side signals high similarity and low variability, and vice versa. A large spread indicates that production variability varies widely across inputs.



(a) Syntactic variability



(b) Syntactic variability

Figure C.2: Human production variability across four NLG tasks. The values on the x-axis are single samples of lexical, syntactic, or semantic similarity between two productions for each input (see Section 10.2). Probability mass on the right side signals high similarity and low variability, and vice versa. A large spread indicates that production variability varies widely across inputs.

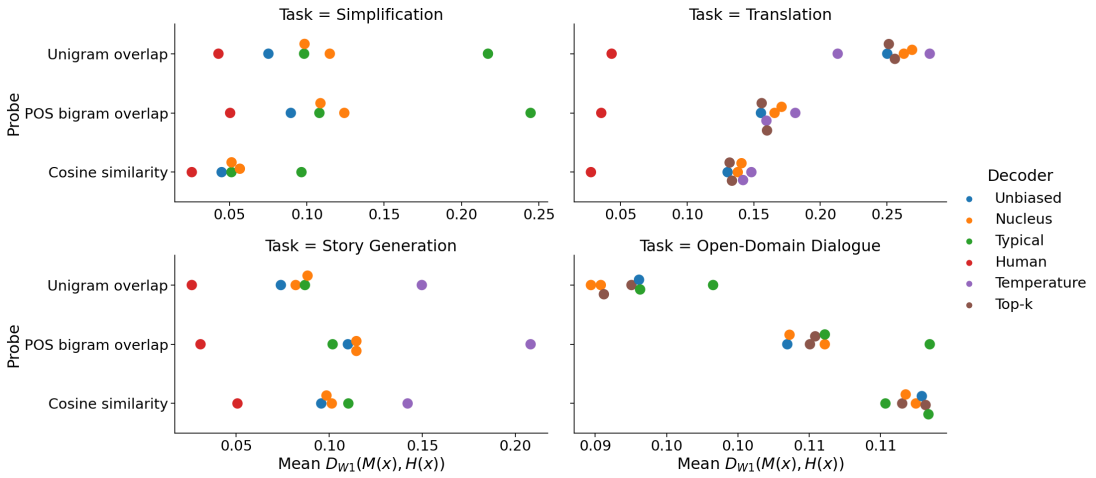


Figure C.3: Mean Wasserstein distances $D_{W_1}(M(x), H(x))$ for (tasks, probe, decoding algorithm) tuples. Base models for each task are described in Section 10.3. Tuples that share colour have different decoding parameters.

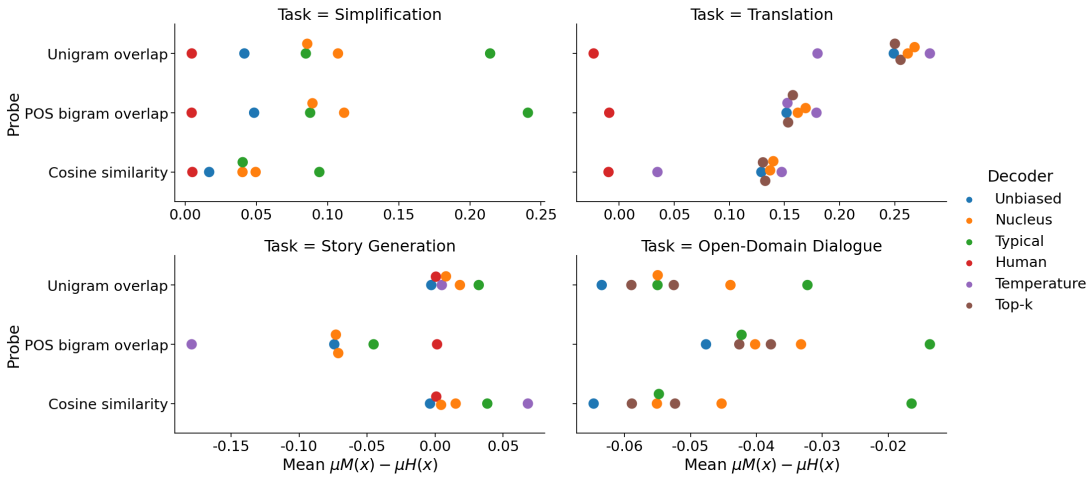


Figure C.4: Mean of distances $\mu_{M(x)} - \mu_{H(x)}$ for (tasks, probe, decoding algorithm) tuples across test sets. Base models for each task are described in Section 10.3. Tuples that share colour have different decoding parameters.

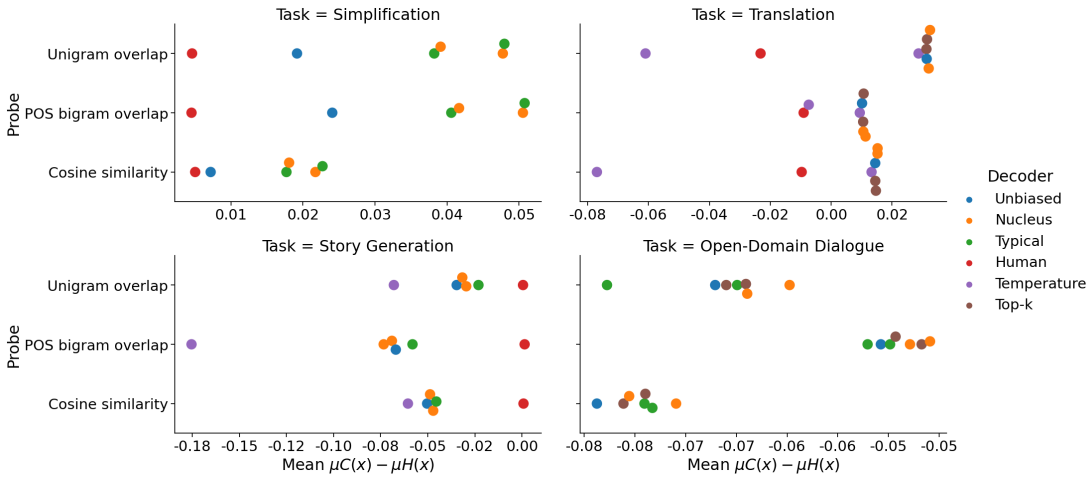


Figure C.5: Mean of distances $\mu_{C(x)} - \mu_{H(x)}$ for (tasks, probe, decoding algorithm) tuples across test sets. Base models for each task are described in Section 10.3. Tuples that share colour have different decoding parameters.

	DailyDialog	**Switchboard**	**WikiText**
GPT-2 Small (124M)	7.34	11.86	25.62
GPT-2 Medium (355M)	6.03	10.50	19.69
GPT-2 Large (774M)	5.26	10.09	17.39
GPT-Neo 125M	7.39	12.54	25.37
GPT-Neo 1.3B	4.94	10.11	14.01
DialoGPT Small	7.94	12.50	-
DialoGPT Medium	6.53	10.96	-
DialoGPT Large	6.23	11.00	-
OPT 125M	17.80	22.68	46.85
OPT 350M	14.88	21.46	40.39
OPT 1.3B	12.58	20.30	27.45

Table C.1: Language model perplexity results. The models tested on the dialogue datasets are fine-tuned for 5 epochs with early stopping; the models tested on WikiText are pre-trained.

C.2 Alternative set generators

For the dialogue corpora, we use GPT-2 (Radford et al., 2019), DialoGPT (Zhang et al., 2020c), and GPT-Neo (Black et al., 2021). For the text corpora, we use GPT-2 (Radford et al., 2019), GPT-Neo (Black et al., 2021), and OPT (Zhang et al., 2022a). The text models are pre-trained while the dialogue models are fine-tuned for 5 epochs with early stopping on the respective datasets, using '`</s> <s>`' as a turn separator. Preliminary experiments on the pre-trained models show that '`</s> <s>`' is the turn separator that yields lowest perplexity on the dialogue datasets. For text models, using no separator is the option that yields the lowest perplexity. When generating out of context, we set x to be either the dialogue turn separator '`</s> <s>`' or a white space for the text models.

LM validation: Perplexity. Table C.1 reports the perplexity of these models on the SWITCHBOARD and DAILYDIALOG test sets, as well as on the WikiText test set (the CLASP dataset and the reading time datasets are too small to allow for fine-tuning, but their style is similar enough to WikiText). Perplexity scores are the lowest for the dialogue datasets. This is to be expected as the dialogue models are fine-tuned. The perplexity of the pre-trained models on WikiText is in line with state-of-the-art results; OPT obtains higher perplexity than GPT-2 and GPT-Neo, but still in an appropriate range.

C.3 Psychometric predictive power and sensitivity of information value estimates

We study the extent to which our estimates of information value are affected by variation in three main factors: the alternative set size ($[10, 20, ..., 100]$), the language model, and the sampling strategy. Figures C.6 and C.7 show Spearman correlation between information value and psychometric data, averaged over subjects. These results complement Sections 11.5.1 and 11.5.2 in Chapter 11.

C.4 Utterance-level surprisal

Given an utterance $\mathbf{y}$ as a sequence of tokens in a context $\mathbf{x}$, token-level surprisal can be defined as $s(y_t) = -\log p(y_t|\mathbf{y}_{<\mathbf{t}}, \mathbf{x})$. Multiple works have proposed quantifying utterance-level surprisal as functions of token-level surprisal (Genzel and Charniak, 2002; Keller, 2004; Xu and Reitter, 2018; Meister et al., 2021; Wallbridge et al., 2022). We compare the predictive power of information value to a number of these utterance-level surprisal aggregates.

Mean surprisal and *total surprisal* account for all token-level surprisal estimates with and without normalising by utterance length:

$$S_{mean}(\mathbf{y}|\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}[s(y_n)] \tag{C.1}$$

$$S_{total}(\mathbf{y}|\mathbf{x}) = \sum_{n=1}^{N}[s(y_n)] \tag{C.2}$$

Superlinear surprisal posits a superlinear effect of token-level estimates:

$$S_{superlinear_k}(\mathbf{y}|\mathbf{x}) = \sum_{n=1}^{N}[s(y_n)]^k \tag{C.3}$$

We experiment with $k \in [0.5, 0.75, \ldots, 5]$. *Maximum surprisal* captures the idea that a highly surprising element drives the overall surprisal of an utterance:

$$S_{max}(\mathbf{y}|\mathbf{x}) = \max[s(y_n)] \tag{C.4}$$

Surprisal variance across an utterance has been defined in a number of ways; we consider *surprisal variance* as the regression to the utterance-level mean and *surprisal difference* as the variability between contiguous token-level estimates:

$$S_{variance}(\mathbf{y}|\mathbf{x}) = \frac{1}{N-1}\sum_{n=2}^{N}[s(y_n) - S_{mean}(\mathbf{y})]^2 \tag{C.5}$$

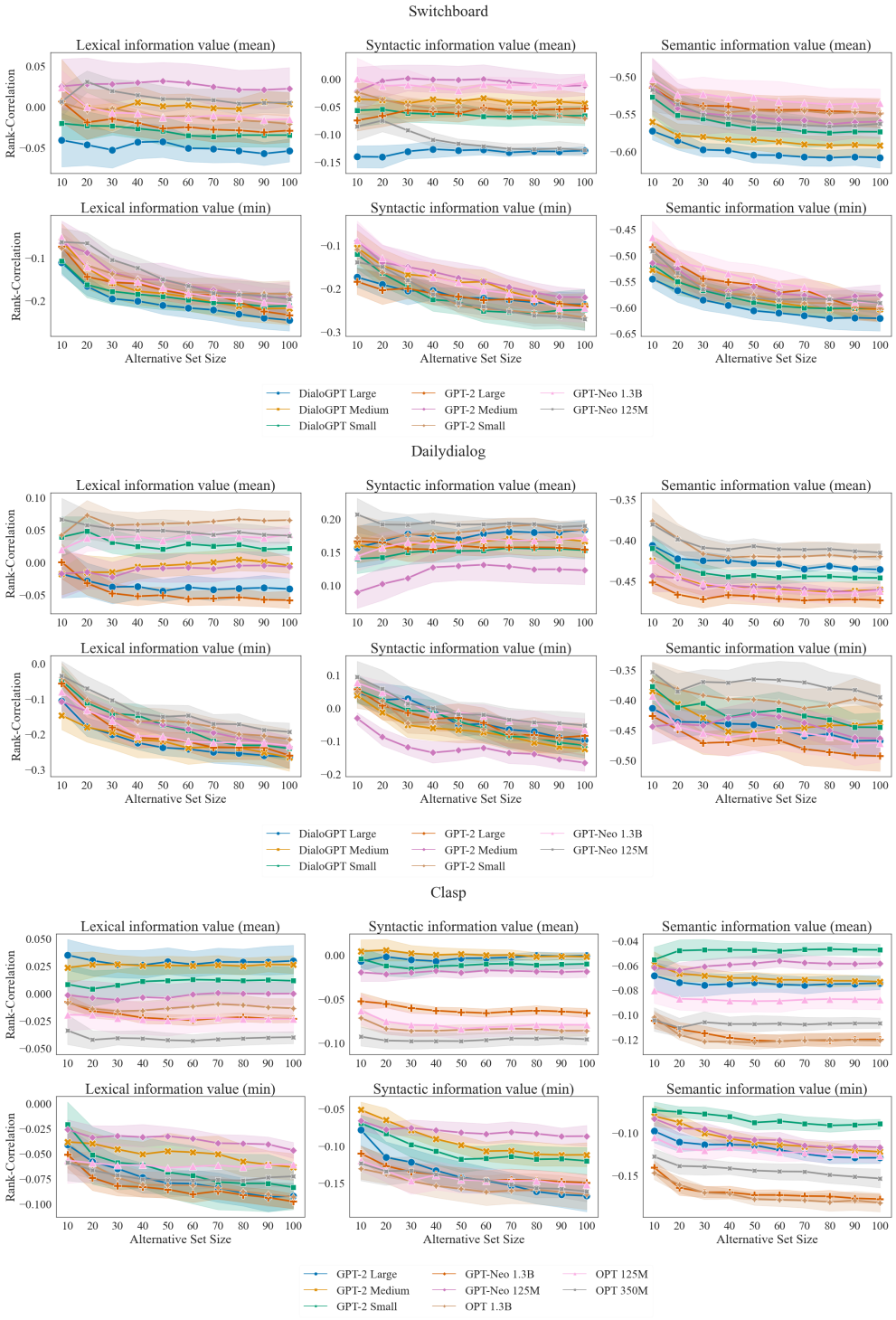


Figure C.6: Spearman correlation between information value and average acceptability judgements.

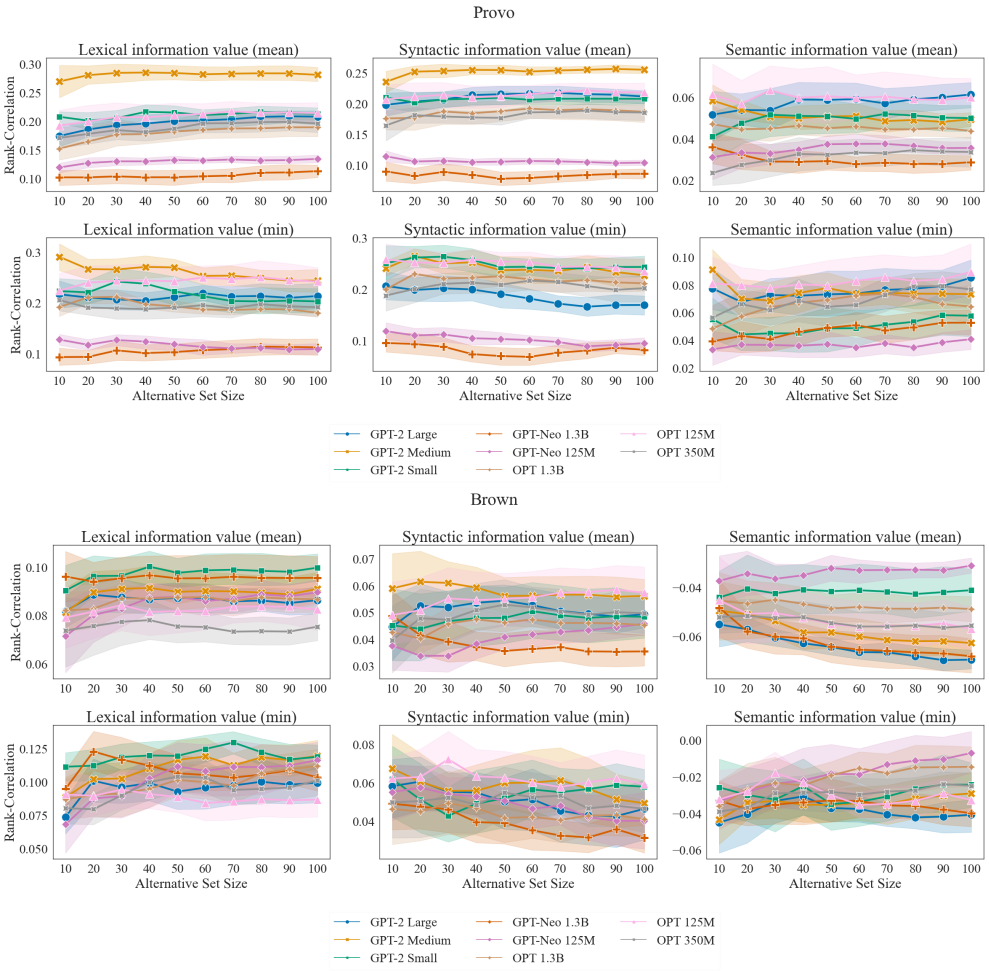


Figure C.7: Spearman correlation between information value and average reading times (length-normalised).

$$S_{difference}(\mathbf{y}|\mathbf{x}) = \sum_{n=2}^{N} |s(y_n) - s(y_{n-1})| \tag{C.6}$$

C.5 Intrinsic robustness analysis

In Section 11.5.1, we evaluate the robustness of information value to parameters involved in the alternative set generation in terms of its psychometric predictive power. We additionally assess their intrinsic robustness by measuring the correlation between information values assigned to target utterances by estimators with different parameter settings.

The parameters which we consider are alternative set size ($[10, 20, ..., 100]$), the generative model, and the decoding strategy. Models and decoding strategies are detailed in Section 11.4.1. For each of the corpora described in Section 11.4.2, we compute the information value for the target utterances based on alternative sets generated under different parameter settings. Robustness is quantified through the distribution of the pairwise Spearman correlation ρ obtained between the information values for each parameter setting; strong pairwise correlation indicates that information value is robust to the varying parameter. Results are displayed in Figures C.8 and C.9 (placed, for convenience, at the end of Appendix C.

Information value defined as lexical, syntactic, and semantic distance becomes highly robust as alternative set size increases; mean correlations between decoding strategies for each model converge towards perfect correlation as alternative set size increases. This pattern holds for all datasets. Decoding strategies do not produce much variation across correlations, regardless of alternative set size (see confidence intervals in Figures C.8 and C.9).

As expected, correlations between parameter settings for min-based distances are more variable. Although they converge to weaker correlations as alternative set size increases when compared to mean-based distances, we still find strong to very strong correlations between decoding strategies for large alternative sets across all models.

C.6 More derived measures of information value

We also tested the following measures derived from information value but found them to be less predictive than those in Chapter 11.

Expected information value. The expected distance of plausible productions given a context x from the alternative set:

$$\mathbb{E}(I(Y|X=x)) \coloneqq \mathbb{E}_{a \in A'_x}\left[I(Y=a, X=x)\right] \tag{C.7}$$

We assume a uniform probability distribution over alternatives. This quantifies the uncertainty over next utterances determined by the context alone. Because the alternative set A_x *is* the set of plausible productions given x, in practice, we compute expected information value using only one alternative set—both in the expectation $\mathbb{E}_{a \in A_x}$ and in the distance calculation $d(y, A_x)$.

Deviation from the expected information value. The absolute difference between the information value for the next utterance y and the expected information value for any next utterance:

$$D(Y{=}y|X{=}x) \coloneqq |I(Y{=}y|X{=}x) - \mathbb{E}(I(Y|X{=}x))| \tag{C.8}$$

This quantifies the information value of an utterance *relative to* the information value expected for plausible productions given x. An analogous notion is the deviation of surprisal from entropy. The token-level version of this forms the basis of the local typicality hypothesis (Meister et al., 2023).

Expected context informativeness. The *expected informativeness* of context x is the reduction in information value contributed by x with respect to any plausible continuation:

$$\mathbb{E}(C(Y{=}y; X{=}x)) \coloneqq \mathbb{E}(I(Y{=}y)) - \mathbb{E}(I(Y{=}y|X{=}x)) \tag{C.9}$$

This quantifies the extent to which a context restricts the space of plausible productions. An analogous notion is the expected pointwise mutual information between $X = x$ and Y, where the value of X is fixed. Similarly to out-of-context information value, out-of-context expected information value $\mathbb{E}(I(Y{=}y))$ is computed with respect to the alternative set $A_{x=\epsilon}$.

C.7 Experimental results: The psychometric predictive power of information value

These are further details about the linear mixed effect models used in Sections 11.6 and 11.7.

Response variables. For PROVO, we use the total dwell time, i.e., the cumulative duration across all fixations on a given word. We filter away any observation that contains 'outlier' words, i.e., words with a z-score > 3 when the distribution of reading times is modelled as log-linear (following Meister et al., 2021).

Control predictors. Following Wilcox et al. (2020), we evaluate each model relative to a baseline model which includes only control variables. Control variables are selected building on previous work (Meister et al., 2021): we include solely an intercept term as a baseline for acceptability judgements and the number of fixated words for reading times. Meister et al. (2021) report similar trends when including summed unigram log probability or sentence length as baseline predictors of acceptability judgements, and word character lengths or word unigram log probabilities for reading times. For reading times, we also test sentence length as a predictor but baseline models that include, instead, the number of fixated words (readers sometimes skip words while reading) achieve higher log-likelihood.

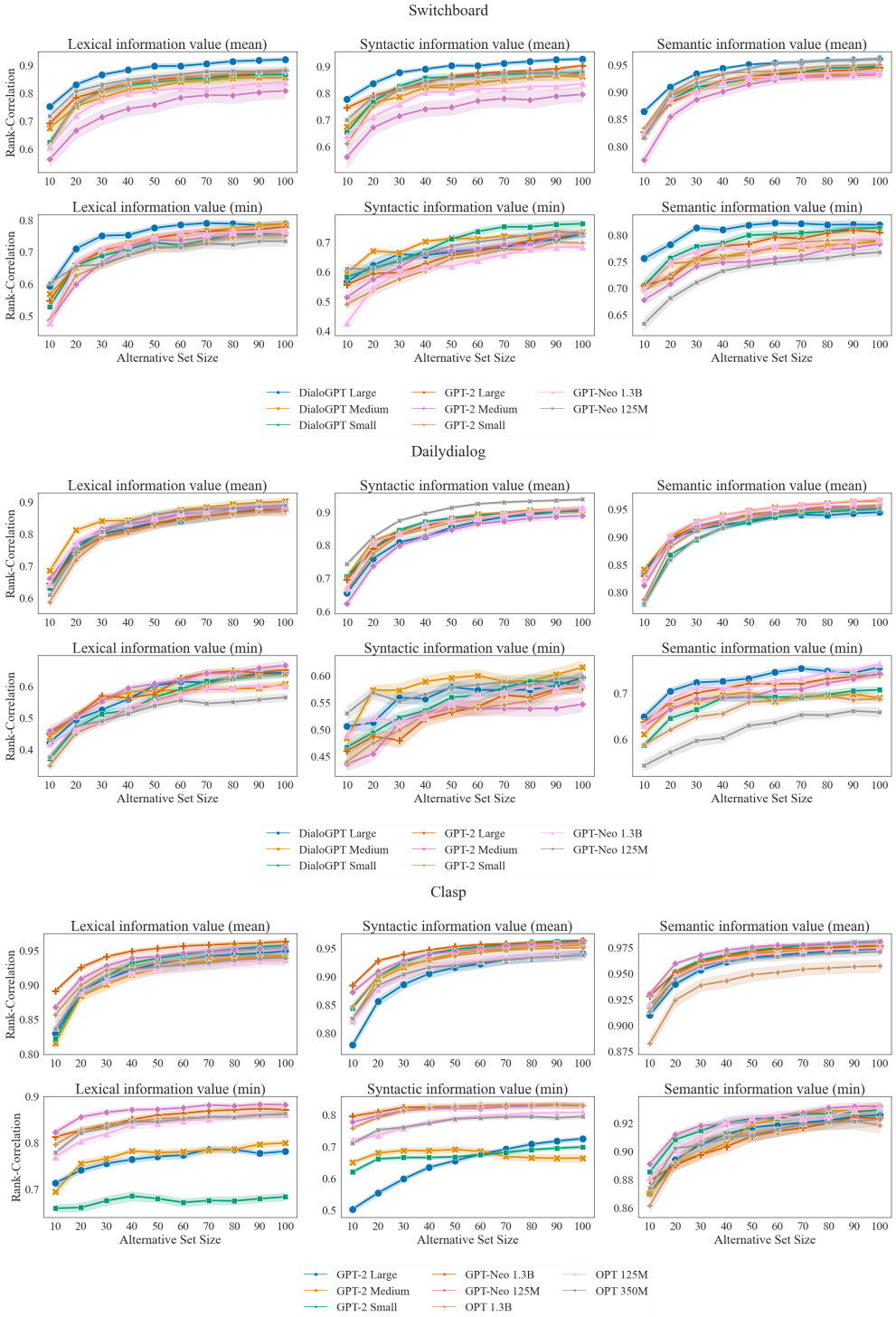


Figure C.8: Intrinsic robustness evaluation on acceptability judgements corpora.

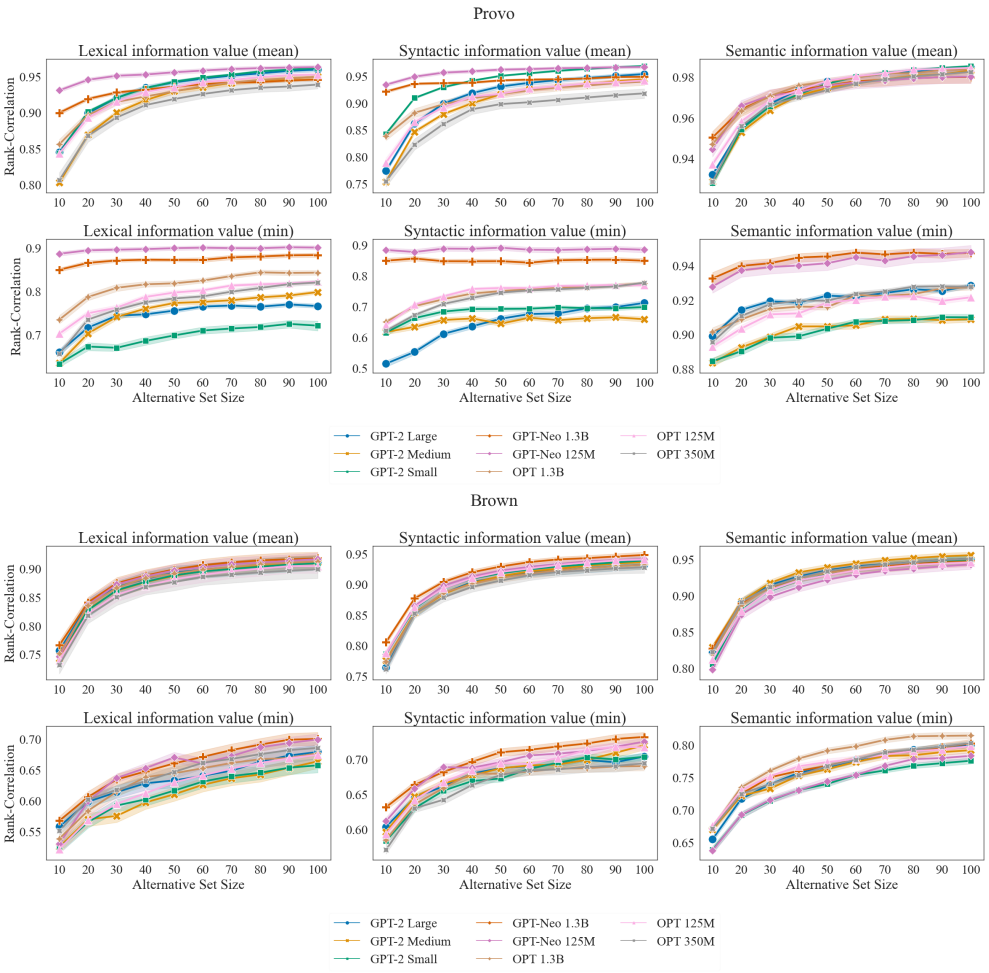


Figure C.9: Intrinsic evaluation on reading times corpora.

Bibliography

L. Aina and T. Linzen. The language model understood the prompt was ambiguous: Probing syntactic uncertainty through generation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 42–57. Association for Computational Linguistics, 2021.

R. Alatrash, D. Schlechtweg, J. Kuhn, and S. Schulte im Walde. CCOHA: Clean corpus of historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966. European Language Resources Association, 2020.

F. Almeman and L. Espinosa Anke. Putting WordNet's dictionary examples in the context of definition modelling: An empirical analysis. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 42–48, 2022.

F. Alva-Manchego, L. Martin, A. Bordes, C. Scarton, B. Sagot, and L. Specia. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679. Association for Computational Linguistics, 2020.

F. Alva-Manchego, C. Scarton, and L. Specia. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4): 861–889, 2021.

S. Amenta, J. Hasenäcker, D. Crepaldi, and M. Marelli. Prediction at the intersection of sentence context and word form: Evidence from eye-movements and self-paced reading. *Psychonomic Bulletin & Review*, 2022.

A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, et al. The HCRC Map Task corpus. *Language and Speech*, 34(4):351–366, 1991.

J. Andreas and D. Klein. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, 2016.

S. Arehalli, B. Dillon, and T. Linzen. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313. Association for Computational Linguistics, 2022.

K. Armeni, R. M. Willems, and S. L. Frank. Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience & Biobehavioral Reviews*, 83:579–588, 2017.

I. Arnon and N. Snider. More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, 62(1):67–82, 2010.

D. Arthur and S. Vassilvitskii. `k-means++`: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

T. August, K. Reinecke, and N. A. Smith. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317. Association for Computational Linguistics, 2022.

J. L. Austin. *How to do things with words*. Oxford University Press, 1975.

M. Aylett and A. Turk. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56, 2004.

M. Aylett and A. Turk. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5):3048–3058, 2006.

J. Baan, W. Aziz, B. Plank, and R. Fernández. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915. Association for Computational Linguistics, 2022.

R. Bamler and S. Mandt. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389, 2017.

M. Banko, V. O. Mittal, and M. J. Witbrock. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 318–325, 2000.

C. Bannard and D. Matthews. Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, 19(3):241–248, 2008.

N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56(1):89–113, 2004.

C. Barkhof and W. Aziz. Statistical model criticism of variational auto-encoders. *arXiv preprint arXiv:2204.03030*, 2022.

M. Baroni. On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic Structures in Natural Language*, pages 1–16. CRC Press, 2022.

M. Baroni and A. Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278, 2013.

D. I. Beaver and B. Z. Clark. *Sense and sensitivity: How focus determines meaning*. John Wiley & Sons, 2009.

A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024, 2003.

A. Belz and A. Gatt. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 197–200. Association for Computational Linguistics, 2008.

L. Bergroth, H. Hakonen, and T. Raita. A survey of longest common subsequence algorithms. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pages 39–48. IEEE, 2000.

J.-P. Bernardy, S. Lappin, and J. H. Lau. The influence of context on sentence acceptability judgements. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461. Association for Computational Linguistics, 2018.

A. Betti and H. Van den Berg. Modelling the history of ideas. *British Journal for the History of Philosophy*, 22(4):812–835, 2014.

M. Bevilacqua, M. Maru, and R. Navigli. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7207–7221. Association for Computational Linguistics, 2020.

D. Biber and F. Barbieri. Lexical bundles in university spoken and written registers. *English for specific purposes*, 26(3):263–286, 2007.

D. Biber, S. Conrad, and V. Cortes. If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3):371–405, 2004.

E. Bigand, B. Tillmann, B. Poulin-Charronnat, and D. Manderlier. Repetition priming: Is music special? *The Quarterly Journal of Experimental Psychology Section A*, 58(8):1347–1375, 2005.

S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

S. Black, G. Leo, P. Wang, C. Leahy, and S. Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, 2021.

A. Blank. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Max Niemeyer Verlag, Berlin, Boston, 1997.

L. Bloomfield. *Language*. Allen & Unwin, 1933.

J. K. Bock. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387, 1986.

O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58. Association for Computational Linguistics, 2014.

G. Boleda. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234, 2020.

G. Boleda and K. Erk. Distributional semantic features as semantic primitives—or not. In *AAAI Spring Symposium Series*, 2015.

S. Borgeaud and G. Emerson. Leveraging sentence similarity in natural language generation: Improving beam search using range voting. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 97–109. Association for Computational Linguistics, 2020.

M. M. Botvinick and D. C. Plaut. Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, 113(2):201, 2006.

S. E. Brennan and H. H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493, 1996.

S. W. Brown. Choosing sense distinctions for WSD: Psycholinguistic evidence. In *Proceedings of ACL-08: HLT, Short Papers*, pages 249–252. Association for Computational Linguistics, 2008.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020a.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020b.

S. Brown-Schmidt, S. O. Yoon, and R. A. Ryskin. People as contexts in conversation. In *Psychology of Learning and Motivation*, volume 62, pages 59–99. Elsevier, 2015.

C. M. Brugman. *The story of Over: Polysemy, semantics, and the structure of the lexicon*. Garland, New York, 1988.

E. Bruni and R. Fernández. Adversarial evaluation for open-domain dialogue generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–288. Association for Computational Linguistics, 2017.

B. Buccola, M. Križ, and E. Chemla. Conceptual alternatives: Competition in language and beyond. *Linguistics and Philosophy*, 45(2):265–291, 2022.

J. Bybee. From usage to grammar: The mind's response to repetition. *Language*, pages 711–733, 2006.

J. Bybee. *Language, usage and cognition*. Cambridge University Press, 2010.

J. Bybee. *Language Change*. Cambridge University Press, 2015.

J. Bybee and J. Scheibman. The effect of usage on degrees of constituency: The reduction of don't in English. *Linguistics*, 37(4):575–596, 1999.

T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27, 1974.

C. Callison-Burch, M. Osborne, and P. Koehn. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256. Association for Computational Linguistics, 2006.

J. Camacho-Collados and M. T. Pilehvar. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63(1):743–788, 2018.

G. Carrol and K. Conklin. Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech*, 63(1):95–122, 2020.

R. Carston. Informativeness, relevance and scalar implicature. *Pragmatics And Beyond New Series*, pages 179–238, 1998.

S. Chen, S. Nathaniel, R. Ryskin, and E. Gibson. The effect of context on noisy-channel sentence comprehension. *Cognition*, 238:105503, 2023.

H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

I. Ciardelli, J. Groenendijk, and F. Roelofsen. *Inquisitive semantics*. Oxford University Press, 2018.

H. H. Clark. *Using Language*. Cambridge University Press, 1996.

H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive Science*, 13 (2):259–294, 1989.

H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986.

M. Clayards, M. K. Tanenhaus, R. N. Aslin, and R. A. Jacobs. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3): 804–809, 2008.

J. D. Cohen, K. Dunbar, and J. L. McClelland. On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, 97(3):332, 1990.

M. X. Collins. Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, 43(5):651–681, 2014.

R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167, 2008.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

G. Columbus. In support of multiword unit classifications: Corpus and human rating data validate phraseological classifications of three different multiword unit types. *Yearbook of Phraseology*, 4(1):23–44, 2013.

K. Conklin and N. Schmitt. The processing of formulaic language. *Annual Review of Applied Linguistics*, 32:45–61, 2012.

L. Connell and D. Lynott. Principles of representation: Why you can't represent the same concept twice. *Topics in Cognitive Science*, 2014.

P. Cook, J. H. Lau, D. McCarthy, and T. Baldwin. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635, 2014.

D. A. Cruse. Polysemy and related phenomena from a cognitive linguistic viewpoint. *Computational lexical semantics*, pages 33–49, 1995.

Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988. Association for Computational Linguistics, 2019.

M. Davies. Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora*, 7(2):121–157, 2012.

F. Davis and M. van Schijndel. Discourse structure interacts with reference but not syntax in neural language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 396–407, 2020.

L. De Mattei, H. Lai, F. Dell'Orletta, and M. Nissim. Human perception in natural language generation. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 15–23. Association for Computational Linguistics, 2021.

F. de Saussure. *Cours de linguistique générale*. Payot, Paris, 1972.

A. de Varda and M. Marelli. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 138–144. Association for Computational Linguistics, 2022.

J. Degen and M. K. Tanenhaus. Processing scalar implicature: A constraint-based approach. *Cognitive Science*, 39(4):667–710, 2015.

J. Degen and M. K. Tanenhaus. Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, 40 (1):172–201, 2016.

G. DeJong. An overview of the FRUMP system. *Strategies for Natural Language Processing*, 113:149–176, 1982.

M. Del Tredici, R. Fernández, and G. Boleda. Short-term meaning shift: A distributional exploration. In *Proceedings of NAACL-HLT 2019 (Annual Conference of the North American Chapter of the Association for Computational Linguistics)*, 2019.

V. Demberg and F. Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210, 2008.

V. Demberg, A. Sayeed, P. Gorinski, and N. Engonopoulos. Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 356–367, 2012.

M. Deng, B. Tan, Z. Liu, E. Xing, and Z. Hu. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605. Association for Computational Linguistics, 2021.

Y. Deng, V. Kuleshov, and A. M. Rush. Model criticism for long-form text generation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

A. Der Kiureghian and O. Ditlevsen. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, 2009.

S. Desai and G. Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 295–302. Association for Computational Linguistics, 2020.

N. Dethlefs, H. Hastie, H. Cuayáhuitl, Y. Yu, V. Rieser, and O. Lemon. Information density and overlap in spoken dialogue. *Computer speech & language*, 37:82–97, 2016.

D. Deutsch, R. Dror, and D. Roth. On the limitations of reference-free evaluations of generated text. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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)*, pages 4171–4186, 2019a.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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)*, pages 4171–4186, 2019b.

Y. Dou, M. Forbes, R. Koncel-Kedziorski, N. A. Smith, and Y. Choi. Is GPT-3 text indistinguishable from human text? Scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274. Association for Computational Linguistics, 2022.

G. Doyle and M. Frank. Shared common ground influences information density in microblog texts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1587–1596. Association for Computational Linguistics, 2015a.

G. Doyle and M. C. Frank. Audience size and contextual effects on information density in Twitter conversations. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 19–28, 2015b.

H. Dubossarsky, Y. Tsvetkov, C. Dyer, and E. Grossman. A bottom up approach to category mapping and meaning change. In *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference*, pages 66–70, 2015.

G. D. Duplessis, F. Charras, V. Letard, A.-L. Ligozat, and S. Rosset. Utterance retrieval based on recurrent surface text patterns. In *European Conference on Information Retrieval*, pages 199–211. Springer, 2017a.

G. D. Duplessis, C. Clavel, and F. Landragin. Automatic measures to characterise verbal alignment in human-agent interaction. In *18th Annual Meeting of the*

Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 71–81, 2017b.

G. D. Duplessis, C. Langlet, C. Clavel, and F. Landragin. Towards alignment strategies in human-agent interactions based on measures of lexical repetitions. *Language Resources and Evaluation*, 55(2):353–388, 2021.

O. Dušek, J. Novikova, and V. Rieser. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Computer Speech & Language*, 59:123–156, 2020.

H. P. Edmundson. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285, 1969.

B. Eikema and W. Aziz. Is MAP decoding all you need? The inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520. International Committee on Computational Linguistics, 2020.

B. Eikema and W. Aziz. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993. Association for Computational Linguistics, 2022a.

B. Eikema and W. Aziz. Sampling-based minimum bayes risk decoding for neural machine translation. In *EMNLP*, 2022b.

J. L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

K. Erk and S. Padó. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906, 2008.

K. Erk and S. Padó. Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference (Short Papers)*, pages 92–97, 2010.

K. Erk, D. McCarthy, and N. Gaylord. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18. Association for Computational Linguistics, 2009.

K. Erk, D. McCarthy, and N. Gaylord. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554, 2013.

M. FAIR Diplomacy Team, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

A. Falaus. Introduction: Alternatives in semantics and pragmatics. *Alternatives in semantics*, pages 1–35, 2013.

I. L. Falkum and A. V. Benito. Polysemy: Current perspectives and approaches. *Lingua: International review of general linguistics*, (157):1–16, 2015.

A. Fan, M. Lewis, and Y. Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898. Association for Computational Linguistics, 2018.

C. Fellbaum. WordNet. In *Theory and applications of ontology: Computer applications*, pages 231–243. Springer, 2010.

P. Fernandes, A. Farinhas, R. Rei, J. G. C. de Souza, P. Ogayo, G. Neubig, and A. Martins. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412. Association for Computational Linguistics, 2022.

J. Firth. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, pages 10–32, 1957.

M. Fomicheva, S. Sun, L. Yankovskaya, F. Blain, F. Guzmán, M. Fishel, N. Aletras, V. Chaudhary, and L. Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555, 2020.

V. Fomin, D. Bakshandaeva, J. Rodina, and A. Kutuzov. Tracing cultural diachronic semantic shifts in Russian using word embeddings: Test sets and baselines. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog Conference*, pages 203–218, 2019.

D. Fox and R. Katzir. On the characterization of alternatives. *Natural language semantics*, 19:87–107, 2011.

A. F. Frank and T. F. Jaeger. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2008.

M. C. Frank and N. D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012a.

M. C. Frank and N. D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012b.

S. L. Frank, L. J. Otten, G. Galli, and G. Vigliocco. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11, 2015. doi: 10.1016/j.bandl.2014.10.006.

M. Franke, G. Jäger, and R. v. Rooij. Vagueness, signaling and bounded rationality. In *JSAI International Symposium on Artificial Intelligence*, pages 45–59. Springer, 2010.

L. Frermann and M. Lapata. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45, 2016.

O. Fugikawa, O. Hayman, R. Liu, L. Yu, T. Brochhagen, and Y. Xu. A computational analysis of crosslinguistic regularity in semantic change. *Frontiers in Communication*, 8:1136338, 2023.

R. Fusaroli, J. Rączaszek-Leonardi, and K. Tylén. Dialog as interpersonal synergy. *New Ideas in Psychology*, 32:147–157, 2014.

R. Futrell and R. Levy. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, long papers*, pages 688–698, 2017.

R. Futrell, E. Wilcox, T. Morita, P. Qian, M. Ballesteros, and R. Levy. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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)*, pages 32–42, 2019.

A. Gadetsky, I. Yakubovskiy, and D. Vetrov. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271. Association for Computational Linguistics, 2018.

W. A. Gale and G. Sampson. Good-Turing frequency estimation without tears. *Journal of quantitative linguistics*, 2(3):217–237, 1995.

N. Gardner, H. Khan, and C.-C. Hung. Definition modeling: Literature review and dataset analysis. *Applied Computing and Intelligence*, 2(1):83–98, 2022.

S. Garrod and A. Anderson. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218, 1987.

G. Gazdar. Pragmatics, implicature, presuposition and logical form. *Critica*, 12 (35), 1979.

D. Geeraerts. *Diachronic Prototype Semantics: A Contribution to Historical Lexicology*. Oxford University Press, 1997.

S. Gehrmann, H. Strobelt, and A. Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116. Association for Computational Linguistics, 2019.

S. Gehrmann, E. Clark, and T. Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*, 2022.

D. Genzel and E. Charniak. Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 199–206, 2002.

D. Genzel and E. Charniak. Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 65–72, 2003.

E. Gibson and J. Thomas. Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14:225–248, 1999.

E. Gibson, L. Bergen, and S. T. Piantadosi. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056, 2013.

T. Glushkova, C. Zerva, R. Rei, and A. F. T. Martins. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938. Association for Computational Linguistics, 2021.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society, 1992.

V. Goel and W. J. Byrne. Minimum Bayes-risk automatic speech recognition. *Computer Speech and Language*, 14(2):115–135, 2000.

A. E. Goldberg. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand, 2006.

A. Goodkind and K. Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18. Association for Computational Linguistics, 2018a.

A. Goodkind and K. Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18. Association for Computational Linguistics, 2018b.

N. D. Goodman and M. C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016.

N. Gotzner and J. Romoli. Meaning and alternatives. *Annual Review of Linguistics*, 8:213–234, 2022.

K. Grewal and Y. Xu. Chaining algorithms and historical adjective extension. In N. Tahmasebi, L. Borin, A. Jatowt, Y. Xu, and S. Hengchen, editors, *Computational approaches to semantic change*, volume 6, page 189. Language Science Press, 2021.

H. P. Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

J. A. G. Groenendijk and M. J. B. Stokhof. *Studies on the Semantics of Questions and the Pragmatics of Answers*. PhD thesis, University of Amsterdam, 1984.

K. Gulordava and M. Baroni. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics, 2011.

J. Haber, T. Baumgärtner, E. Takmaz, L. Gelderloos, E. Bruni, and R. Fernández. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, 2019.

J. Hale. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001.

J. Y. Halpern. *Reasoning about uncertainty*. MIT press, 2017.

C. L. Hamblin. Questions in Montague English. In *Montague grammar*, pages 247–259. Elsevier, 1976.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, 2016.

Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

T. B. Hashimoto, H. Zhang, and P. Liang. Unifying human and statistical evaluation for natural language generation. 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)*, pages 1689–1701. Association for Computational Linguistics, 2019.

R. Hawkins, M. Kwon, D. Sadigh, and N. Goodman. Continual adaptation for efficient machine communication. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 408–419, 2020a.

R. D. Hawkins, M. C. Frank, and N. D. Goodman. Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, 44(6):e12845, 2020b.

M. Heilman, A. Cahill, N. Madnani, M. Lopez, M. Mulholland, and J. Tetreault. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180. Association for Computational Linguistics, 2014.

D. S. Hirschberg. Algorithms for the longest common subsequence problem. *Journal of the ACM (JACM)*, 24(4):664–675, 1977.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

P. Hofmeister, L. S. Casasanto, and I. A. Sag. Processing effects in linguistic judgment data: (Super-)additivity and reading span scores. *Language and Cognition*, 6:111 – 145, 2014.

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*, 2019.

A. Holtzman, P. West, V. Shwartz, Y. Choi, and L. Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051. Association for Computational Linguistics, 2021.

M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. spaCy: Industrial-strength natural language processing in Python. 2020. doi: 10.5281/zenodo.1212303.

P. J. Hopper et al. On some principles of grammaticization. *Approaches to Grammaticalization*, 1:17–35, 1991.

L. R. Horn. *On the semantic properties of logical operators in English*. University of California, Los Angeles, 1972.

L. R. Horn and G. L. Ward. *The Handbook of Pragmatics*. Wiley Online Library, 2004.

J. Hu, R. Levy, and S. Schuster. Predicting scalar diversity with context-driven uncertainty over alternatives. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 68–74. Association for Computational Linguistics, 2022.

J. Hu, R. Levy, J. Degen, and S. Schuster. Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics*, 2023. To appear.

R. Hu, S. Li, and S. Liang. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908. Association for Computational Linguistics, 2019.

H. Huang, T. Kajiwara, and Y. Arase. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509. Association for Computational Linguistics, 2021.

J. Huang, H. Shao, K. C.-C. Chang, J. Xiong, and W.-m. Hwu. Understanding jargon: Combining extraction and generation for definition modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4004. Association for Computational Linguistics, 2022.

E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

D. Hupkes, M. Giulianelli, V. Dankers, M. Artetxe, Y. Elazar, T. Pimentel, C. Christodoulopoulos, K. Lasri, N. Saphra, A. Sinclair, et al. A taxonomy and review of generalisation research in NLP. *Nature Machine Intelligence*, 2023.

S. Ishiwatari, H. Hayashi, N. Yoshinaga, G. Neubig, S. Sato, M. Toyoda, and M. Kitsuregawa. Learning to describe unknown phrases with local and global contexts. 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), pages 3467–3476. Association for Computational Linguistics, 2019.

T. F. Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62, 2010.

T. F. Jaeger and R. P. Levy. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systemss*, pages 849–856, 2007.

T. F. Jaeger and H. Tily. On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):323–335, 2011.

N. Janssen and H. A. Barber. Phrase frequency effects in language production. *PloS one*, 7(3):e33202, 2012.

F. Jelinek and R. L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proceeding of the Workshop on Pattern Recognition in Practice*, pages 381–397, 1980.

F. Jelinek, L. Bahl, and R. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21(3):250–256, 1975.

P. N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, USA, 1986.

K. Jokinen. Goal formulation based on communicative principles. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, 1996.

H. Jolsvai, S. M. McCauley, and M. H. Christiansen. Meaning overrides frequency in idiomatic and compositional multiword chunks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2013.

D. Jurafsky and J. H. Martin. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA, 2009.

A. Kapron-King and Y. Xu. A diachronic evaluation of gender asymmetry in euphemism. *LChange'21*, 2021:28–38, 2021.

S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401, 1987.

R. Katzir. Structurally-defined alternatives. *Linguistics and Philosophy*, 30:669–690, 2007.

F. Keller. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 317–324, 2004.

A. Kilgarriff. I don't believe in word senses. *Computers and the Humanities*, 31 (2):91–113, 1997.

Y. Kim, Y.-I. Chiu, K. Hanaki, D. Hegde, and S. Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, 2014.

W. Kintsch. Meaning in context. In *Handbook of Latent Semantic Analysis*, pages 101–118. Psychology Press, 2007.

D. E. Klein and G. L. Murphy. The representation of polysemous words. *Journal of Memory and Language*, 45(2):259–282, 2001.

R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184. IEEE, 1995.

D. Koller and N. Friedman. *Probabilistic graphical models: Principles and techniques*. MIT press, 2009.

C. Kong, Y. Chen, H. Zhang, L. Yang, and E. Yang. Multitasking framework for unsupervised simple definition generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5934–5943. Association for Computational Linguistics, 2022.

E. Krahmer and K. van Deemter. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218, 2012.

B. Krause, E. Kahembwe, I. Murray, and S. Renals. Dynamic evaluation of neural sequence models. In *International Conference on Machine Learning*, pages 2766–2775. PMLR, 2018.

R. M. Krauss and S. Weinheimer. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1(1):113–114, 1964.

R. M. Krauss and S. Weinheimer. Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning and Verbal Behavior*, 6 (3):359–363, 1967.

J. K. Kruschke. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1):22, 1992.

L. Kuhn, Y. Gal, and S. Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.

K. Kuiper. *Smooth talkers: The linguistic performance of auctioneers and sportscasters*. Routledge, 1995.

V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee, 2015.

S. Kumar and B. Byrne. Minimum Bayes-risk word alignments of bilingual texts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 140–147, 2002.

S. Kumar and W. Byrne. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, 2004.

S. Kurtyigit, M. Park, D. Schlechtweg, J. Kuhn, and S. Schulte im Walde. Lexical semantic change discovery. 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)*, pages 6985–6998. Association for Computational Linguistics, 2021.

A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, 2018.

B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

J. H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601. Association for Computational Linguistics, 2012.

J. H. Lau, A. Clark, and S. Lappin. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1618–1628, 2015.

J. H. Lau, A. Clark, and S. Lappin. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5): 1202–1241, 2017.

S. Lawrence, C. L. Giles, and S. Fong. Natural language grammatical inference with recurrent neural networks. *IEEE Transactions on Knowledge & Data Engineering*, 12(01):126–140, 2000.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

D. Lee, K. C. Cheung, and N. L. Zhang. Adaptive label smoothing with self-knowledge in natural language generation. *arXiv preprint arXiv:2210.13459*, 2022.

W. J. Levelt. *Speaking: From intention to articulation*, volume 1. MIT press, 1993.

S. C. Levinson. *Pragmatics*. Cambridge University Press, 1983.

S. C. Levinson, C. Stephen, and S. C. Levinson. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press, 2000.

R. Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008a.

R. Levy. A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 234–243, 2008b.

R. Levy and T. F. Jaeger. Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

R. Levy, K. Bicknell, T. Slattery, and K. Rayner. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50):21086–21090, 2009.

R. P. Levy. Communicative efficiency, uniform information density, and the rational speech act theory. In *40th Annual Meeting of the Cognitive Science Society*, pages 684–689. Cognitive Science Society, 2018.

J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, 2016.

Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995. Asian Federation of Natural Language Processing, 2017.

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics, 2004.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132. Association for Computational Linguistics, 2016.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

R. Love, C. Dembry, A. Hardie, V. Brezina, and T. McEnery. The Spoken BNC 2014. *International Journal of Corpus Linguistics*, 22(3):319–344, 2017.

R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126. Association for Computational Linguistics, 2017.

L. Lucy and D. Bamman. Characterizing English variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556, 2021.

L. Lucy, D. Tadimeti, and D. Bamman. Discovering differences in the representation of people using contextualized semantic axes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3477–3494, 2022.

P. Ludlow. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. Oxford University Press, 2014.

H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

S. Luke and K. Christianson. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833, 2018.

W. J. Ma and B. Peters. A neural network walks into a lab: Towards using deep nets as models for human behavior. *arXiv preprint arXiv:2005.02181*, 2020.

A. Malinin and M. Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2020.

S. Manandhar, I. P. Klapaftis, D. Dligach, and S. S. Pradhan. SemEval-2010 Task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68. Association for Computational Linguistics, 2010.

N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220, 1967.

M. Martinc, S. Montariol, E. Zosa, and L. Pivovarova. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the International World Wide Web Conference*, pages 20–24, 2020.

B. McCann, J. Bradbury, C. Xiong, and R. Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305, 2017.

D. McCarthy. Word sense disambiguation: An overview. *Language and Linguistics compass*, 3(2):537–558, 2009.

J. L. McClelland and T. T. Rogers. The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4(4):310–322, 2003.

J. L. McClelland, D. E. Rumelhart, P. R. Group, et al. Parallel distributed processing. *Explorations in the Microstructure of Cognition*, 2:216–271, 1986.

J. L. McClelland, F. Hill, M. Rudolph, J. Baldridge, and H. Schütze. Extending machine language models toward human-level language understanding. *arXiv preprint arXiv:1912.05877*, 2019.

D. D. McDonald. Subsequent reference: Syntactic and rhetorical constraints. In *Proceedings of the 1978 Workshop on Theoretical issues in natural language processing*, pages 64–72, 1978.

S. A. McDonald and R. C. Shillcock. Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research*, 43(16):1735–1751, 2003.

C. Meister and R. Cotterell. Language model evaluation beyond perplexity. 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)*, pages 5328–5339. Association for Computational Linguistics, 2021.

C. Meister, R. Cotterell, and T. Vieira. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2185. Association for Computational Linguistics, 2020.

C. Meister, T. Pimentel, P. Haller, L. Jäger, R. Cotterell, and R. Levy. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980. Association for Computational Linguistics, 2021.

C. Meister, T. Pimentel, G. Wiher, and R. Cotterell. Locally Typical Sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023.

D. Merkx and S. L. Frank. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, 2021.

T. Mickus, D. Paperno, and M. Constant. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 1–11. Linköping University Electronic Press, 2019.

T. Mickus, K. Van Deemter, M. Constant, and D. Paperno. SemEval-2022 Task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14. Association for Computational Linguistics, 2022.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993.

M. P. Mitchell, B. Santorini, M. A. Marcinkiewicz, and A. Taylor. Treebank-3 LDC99T42 Web Download. Linguistic Data Consortium, 1999.

S. Mitra, R. Mitra, M. Riedl, C. Biemann, A. Mukherjee, and P. Goyal. That's sick dude! Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029, 2014.

W. Monroe, R. X. Hawkins, N. D. Goodman, and C. Potts. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338, 2017.

I. F. Monsalve, S. L. Frank, and G. Vigliocco. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408, 2012.

S. Montariol, M. Martinc, and L. Pivovarova. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, 2021.

R. Navigli. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69, 2009.

R. Navigli and D. Vannella. SemEval-2013 Task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, 2013.

B. Newman, R. Cohn-Gordon, and C. Potts. Communication-based evaluation for natural language generation. *Proceedings of the Society for Computation in Linguistics*, 3(1):234–244, 2020.

K. Ni and W. Y. Wang. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417. Asian Federation of Natural Language Processing, 2017.

B. Noble, A. Sayeed, R. Fernández, and S. Larsson. Semantic shift in social networks. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 26–37. Association for Computational Linguistics, 2021.

T. Noraset, C. Liang, L. Birnbaum, and D. Downey. Definition modeling: Learning to define word embeddings in natural language. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

B.-D. Oh and W. Schuler. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *arXiv preprint arXiv:2212.12131*, 2022.

M. Ott, M. Auli, D. Grangier, and M. Ranzato. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*, pages 3956–3965. PMLR, 2018.

P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 613–619. Association for Computing Machinery, 2002.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

C. Paradis. Metonymization: A key mechanism in semantic change. *Defining Metonymy in Cognitive Linguistics: Towards a Consensus View*, pages 61–98, 2011.

H. Paul. *Prinzipien der Sprachgeschichte*. Halle: Max Niemeyer, 1886.

C. S. Peirce. Collected papers. Ed. Charles Hartshorne and Paul Weiss. *Cambridge, MA: Harvard University Press*, 2:16, 1932.

J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

M. Peters, W. Ammar, C. Bhagavatula, and R. Power. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, 2017.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

M. J. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–190, 2004.

M. T. Pilehvar and J. Camacho-Collados. WiC: The Word-in-Context dataset for evaluating context-sensitive meaning representations. 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)*, pages 1267–1273. Association for Computational Linguistics, 2019.

K. Pillutla, S. Swayamdipta, R. Zellers, J. Thickstun, S. Welleck, Y. Choi, and Z. Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc., 2021.

T. Pimentel, C. Meister, E. Salesky, S. Teufel, D. Blasi, and R. Cotterell. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2021.

T. Pimentel, C. Meister, and R. Cotterell. Cluster-based evaluation of automatically generated text. *arXiv preprint arXiv:2205.16001*, 2022.

D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

T. Qian and T. F. Jaeger. Topic shift in efficient discourse production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2011.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

M. A. Ré and R. K. Azad. Generalization of entropy based divergence measures for symbolic sequence analysis. *PloS one*, 9(4):e93532, 2014.

R. Reddy. Speech understanding systems: A summary of results of the five-year research effort at Carnegie Mellon University. Technical report, 1977.

R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702. Association for Computational Linguistics, 2020.

R. Rei, A. C. Farinha, C. Zerva, D. van Stigt, C. Stewart, P. Ramos, T. Glushkova, A. F. T. Martins, and A. Lavie. Are references really needed? Unbabel-IST 2021 submission for the Metrics Shared Task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040. Association for Computational Linguistics, 2021.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019.

E. Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, 2018.

E. Reiter and R. Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

D. Reitter and J. D. Moore. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46, 2014.

D. Reitter, F. Keller, and J. D. Moore. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, companion volume: Short papers*, pages 121–124, 2006.

T. T. Rogers and J. L. McClelland. *Semantic cognition: A parallel distributed processing approach*. MIT press, 2004.

M. Rooth. A theory of focus interpretation. *Natural language semantics*, pages 75–116, 1992.

M. Rooth. Focus. The handbook of contemporary semantic theory, ed. by Shalom Lappin, 1996.

S. Rosenberg and B. D. Cohen. Speakers' and listeners' processes in a word-communication task. *Science*, 145(3637):1201–1203, 1964.

A. Rosenfeld and K. Erk. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484, 2018.

P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

M. Rudolph and D. Blei. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1003–1011. International World Wide Web Conferences Steering Committee, 2018.

D. E. Rumelhart and J. L. McClelland. *On Learning the Past Tenses of English Verbs*, pages 216–271. MIT Press, Cambridge, MA, USA, 1986.

D. E. Rumelhart, G. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

H. Sacks. Lectures on conversation: Volume I. *Malden, Massachusetts: Blackwell*, 1992.

H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.

A. B. Sai, A. K. Mohankumar, S. Arora, and M. M. Khapra. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827, 2020.

P. Sander. Beyond broadening and narrowing: Investigating the nature and consequences of change in lexical semantic change detection. *Unpublished master's thesis, Saarland University. Saarbrücken, Germany*, 2023.

A. J. Sanford and S. C. Garrod. *Understanding written language: Explorations of comprehension beyond the sentence*. John Wiley & Sons, 1981.

C. Scheepers. Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, 89(3):179–205, 2003.

E. A. Schegloff. On talk and its institutional occasions. *Talk at work: Interaction in institutional settings*, pages 101–134, 1992.

T. Schick, J. Dwivedi-Yu, Z. Jiang, F. Petroni, P. Lewis, G. Izacard, Q. You, C. Nalmpantis, E. Grave, and S. Riedel. PEER: A collaborative language model. *arXiv preprint arXiv:2208.11663*, 2022.

D. Schlechtweg, S. Schulte im Walde, and S. Eckmann. Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In

Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), volume 2, pages 169–174, 2018.

D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, and N. Tahmasebi. SemEval-2020 Task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, 2020.

D. Schlechtweg, N. Tahmasebi, S. Hengchen, H. Dubossarsky, and B. McGillivray. DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091. Association for Computational Linguistics, 2021.

M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), 2021. doi: 10.1073/pnas.2105646118.

H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24 (1):97–123, 1998.

J. R. Searle. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge University Press, 1969.

A. See, A. Pappu, R. Saxena, A. Yerukola, and C. D. Manning. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pages 843–861. Association for Computational Linguistics, 2019.

P. Seedhouse. Conversation analysis. In R. Bayley, R. Cameron, and C. Lucas, editors, *The Oxford Handbook of Sociolinguistics*. Oxford University Press, 2013.

T. Sellam, D. Das, and A. Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. Association for Computational Linguistics, 2020.

C. Shain, I. A. Blank, M. van Schijndel, W. Schuler, and E. Fedorenko. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307, 2020.

C. Shain, C. Meister, T. Pimentel, R. Cotterell, and R. P. Levy. Large-scale evidence for logarithmic effects of word predictability on reading time. *PsyArXiv preprint*, 2022.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

T. Shore, T. Androulakaki, and G. Skantze. KTH tangrams: A dataset for research on alignment and conceptual pacts in task-oriented dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 2018.

H. A. Simon. Bounded rationality. In *Utility and probability*, pages 15–18. Springer, 1990.

A. Sinclair and R. Fernández. Construction coordination in first and second language acquisition. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*. SEMDIAL, 2021.

A. Sinclair, J. Jumelet, W. Zuidema, and R. Fernández. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050, 2022.

K. Sinha, P. Parthasarathi, J. Wang, R. Lowe, W. L. Hamilton, and J. Pineau. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441. Association for Computational Linguistics, 2020.

A. Siyanova-Chanturia, K. Conklin, and W. J. Van Heuven. Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3):776, 2011.

A. Siyanova-Chanturia, K. Conklin, S. Caffarra, E. Kaan, and W. J. van Heuven. Representation and processing of multi-word expressions in the brain. *Brain and language*, 175:111–122, 2017.

N. J. Smith and R. Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319, 2013.

H. Somers. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 127–133, 2005.

P. Sommerauer and A. Fokkens. Conceptual change and distributional semantic models: An exploratory study on pitfalls and possibilities. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 223–233, 2019.

A. Sorace and F. Keller. Gradience in linguistic data. *Lingua*, 115(11):1497–1524, 2005.

D. Sperber and D. Wilson. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA, 1986.

R. C. Stalnaker. Assertion. In *Pragmatics*, pages 315–332. Brill, 1978.

K. Stasaski and M. A. Hearst. Pragmatically appropriate diversity for dialogue evaluation. *arXiv preprint arXiv:2304.02812*, 2023.

M. Straka and J. Straková. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99. Association for Computational Linguistics, 2017.

T. Sumers, R. Hawkins, M. K. Ho, and T. Griffiths. Extending rational models of communication from beliefs to actions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.

M. Suzgun, L. Melas-Kyriazi, and D. Jurafsky. Follow the wisdom of the crowd: Effective text generation via minimum bayes risk decoding. *arXiv preprint arXiv:2211.07634*, 2022.

T. Szymanski. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453. Association for Computational Linguistics, 2017.

P. Tabossi, R. Fanari, and K. Wolf. Why are idioms recognized fast? *Memory & Cognition*, 37(4):529–540, 2009.

N. Tahmasebi, L. Borin, and A. Jatowt. Survey of computational approaches to diachronic conceptual change detection. *Computational Linguistics*, 1(1), 2018.

N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky, and L. Borin, editors. *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 2022. Association for Computational Linguistics.

E. Takmaz, M. Giulianelli, S. Pezzelle, A. Sinclair, and R. Fernández. Refer, reuse, reduce: Generating subsequent references in visual and conversational contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4350–4368. Association for Computational Linguistics, 2020.

X. Tang, W. Qu, and X. Chen. Semantic change computation: A successive approach. *World Wide Web*, 19(3):375–415, 2016.

R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

J. Tiedemann and S. Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, 2020.

D. Titone and M. Libben. Time-dependent effects of decomposability, familiarity and literal plausibility on idiom priming: A cross-modal priming investigation. *The Mental Lexicon*, 9(3):473–496, 2014.

D. A. Titone and C. M. Connine. Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbol*, 9(4):247–270, 1994.

M. Tomasello. *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press, 2003.

E. C. Traugott. Semantic change. In *Oxford Research Encyclopedia of Linguistics*. 2017.

A. Tremblay and R. H. Baayen. Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. *Perspectives on formulaic language: Acquisition and communication*, pages 151–173, 2010.

A. Tremblay, B. Derwing, G. Libben, and C. Westbury. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language learning*, 61(2):569–613, 2011.

J. Turian, L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

G. Underwood, N. Schmitt, and A. Galpin. The eyes have it: An eye movement study into the processing of formulaic sequences. In N. Schmitt, editor, *Formulaic Sequences: Acquisition, Processing and Use*, pages 153–172. John Benjamins, 2004.

T. A. Van Dijk, W. Kintsch, et al. *Strategies of discourse comprehension.* Academic Press, 1983.

M. van Schijndel and T. Linzen. A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710, 2018.

B. Van Tiel, E. Van Miltenburg, N. Zevakhina, and B. Geurts. Scalar Diversity. *Journal of Semantics*, 33(1):137–175, 2014.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

A. Vega and N. Ward. Looking for entropy rate constancy in spoken dialog. Technical Report UTEP-CS-09-19, University of Texas El Paso, 2009.

V. Verma, N. Tomlin, and D. Klein. Revisiting entropy rate constancy in text. *arXiv preprint arXiv:2305.12084*, 2023.

M. Wagner et al. *Prosody and recursion.* PhD thesis, Massachusetts Institute of Technology, 2005.

S. Wallbridge, P. Bell, and C. Lai. Investigating perception of spoken dialogue acceptability through surprisal. In *Interspeech 2022: The 23rd Annual Conference of the International Speech Communication Association*, pages 4506–4510. International Speech Communication Association, 2022.

S. Wallbridge, P. Bell, and C. Lai. Do dialogue representations align with perception? an empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2696–2713. Association for Computational Linguistics, 2023.

A. Warstadt, A. Singh, and S. R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

J. Wei, C. Meister, and R. Cotterell. A cognitive regularizer for language modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 5191–5202. Association for Computational Linguistics, 2021.

G. Wiedemann, S. Remus, A. Chawla, and C. Biemann. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019*, 2019.

D. T. Wijaya and R. Yeniterzi. Understanding semantic change of words over centuries. In *Proceedings of the 2011 International Workshop on Detecting and Exploiting Cultural Diversity on the Social Web*, pages 35–40. ACM, 2011.

E. G. Wilcox, J. Gauthier, J. Hu, P. Qian, and R. Levy. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713. Cognitive Science Society, 2020.

E. G. Wilcox, T. Pimentel, C. Meister, R. Cotterell, and R. P. Levy. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 2023.

R. M. Willems, S. L. Frank, A. D. Nijhof, P. Hagoort, and A. Van den Bosch. Prediction during natural language comprehension. *Cerebral Cortex*, 26(6): 2506–2516, 2016.

L. Wittgenstein. Philosophische Untersuchungen. *Macmillan*, 1953.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, 2020.

A. Wray. *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press, 2002.

W. Xu, C. Callison-Burch, and C. Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297, 2015.

W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016.

Y. Xu and C. Kemp. A computational evaluation of two laws of semantic change. In *Proceedings of CogSci*, 2015.

Y. Xu and D. Reitter. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170:147–163, 2018.

S. Yan and T. F. Jaeger. Expectation adaptation during natural reading. *Language, Cognition and Neuroscience*, 35(10):1394–1422, 2020.

W. Yuan, G. Neubig, and P. Liu. BARTScore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc., 2021.

F. D. Zamora-Reina, F. Bravo-Marquez, and D. Schlechtweg. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164. Association for Computational Linguistics, 2022.

N. Zaslavsky, J. Hu, and R. P. Levy. A rate–distortion view of human pragmatic reasoning? In *Proceedings of the Society for Computation in Linguistics 2021*, pages 347–348. Association for Computational Linguistics, 2021.

S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022a.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020a.

X. Zhang, Y. Liu, X. Wang, P. He, Y. Yu, S.-Q. Chen, W. Xiong, and F. Wei. Momentum calibration for text generation. *arXiv preprint arXiv:2212.04257*, 2022b.

Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. DialoGPT: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*, 2020b.

Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and W. B. Dolan. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 2020c.

Z. Zhang, L. Bergen, A. Paunov, R. Ryskin, and E. Gibson. Scalar implicature is sensitive to contextual alternatives. *Cognitive Science*, 47(2):e13238, 2023.

Y. Zhao, M. Khalman, R. Joshi, S. Narayan, M. Saleh, and P. J. Liu. Calibrating sequence likelihood improves conditional language generation. *arXiv preprint arXiv:2210.00045*, 2022.

M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*, 2022.

Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27, 2015.

G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, 1949.

R. A. Zwaan and G. A. Radvansky. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162, 1998.

Samenvatting

Dit proefschrift onderzoekt nieuwe manieren om kunstmatige neurale netwerken te gebruiken als modellen van menselijk taalgebruik, met als doel nieuwe methoden te ontwikkelen en nieuwe onderzoeksrichtingen mogelijk te maken voor een grote verscheidenheid aan taalwetenschappers: van historische taalkundigen, sociolinguïsten en lexicografen tot cognitieve wetenschappers en neurowetenschappers. Het bestaat uit een reeks studies over taalbegrip en taalproductie, met de nadruk op hoe hun modellering wordt beïnvloed wanneer taalkundige contexten op passende wijze in overweging worden genomen.

Deel 1 presenteert twee nieuwe methoden om woordgebruik te bestuderen als functie van de zinscontext waarin een woord voorkomt: de eerste bestaat uit het extraheren, groeperen en analyseren van gecontextualiseerde neurale representaties uit taalmodellen; de tweede maakt gebruik van door taalmodellen gegenereerde menselijk leesbare woorddefinities op basis van voorbeelden van woordgebruik. Lexicale semantische veranderingsanalyse wordt hier als voorbeeld genomen, aangezien dit het dynamisch vastleggen van woordbetekenis vereist vanwege zijn genuanceerde contextbepaalde modulaties.

Deel 2 richt zich op neurale modellen als contextbewuste simulaties van taal ontvangers. Ik verkrijg 'surprisal' schattingen van de voorspelbaarheid van uitingen vanuit neurale taalmodellen en gebruik deze om psycholinguïstische theorieën over de productie van uitingen te testen, die het monitoren van de voorspelbaarheid van sprekers postuleren. Bevindingen dagen gevestigde hypothesen van rationeel gebruik van het communicatiekanaal uit, vooral in dialogische omgevingen—maar over het algemeen bevestigen ze dat de strategieën voor het produceren van uitingen kunnen worden beschreven als het efficiënt beperken van de inspanning die het gesprekspartners vergt om de boodschap te begrijpen.

Deel 3 onderzoekt het potentieel van neurale tekstgeneratoren als modellen van taalproductie. Ik test of generatoren taal produceren dat statistische eigenschappen bevat die overeenkomen met die van menselijke producties, en gebruik ze vervolgens om interpreteerbare metingen van voorspelbaarheid van uitingen te

verkrijgen die complementair zijn aan de maten die in deel 2 zijn gebruikt. Ik sluit af door inzichten uit de rest van dit proefschrift te verzamelen in een formeel kader voor kunstmatige simulaties van mensachtig—efficiënt, communicatief effectief en publiek bewust—taalproductie gedrag.

Abstract

This thesis explores novel ways of using artificial neural networks as models of human language use, with the goal of establishing new methods and enabling new research directions for a wide variety of language scientists: from historical linguists, sociolinguists, and lexicographers to cognitive scientists and neuroscientists. It consists of a series of studies on language comprehension and language production. with an emphasis on how their modelling is affected when linguistic contexts are properly taken into account. It is divided into three parts.

Part 1 presents two novel methods to study word usage as a function of a word's sentential context of occurrence: the first consists of extracting, grouping, and analysing contextualised neural representations from language models; the second uses human-readable word definitions generated by language models prompted with word usage examples. Lexical semantic change analysis is taken as an example application, as it requires dynamically capturing word meaning with its nuanced context-determined modulations.

Part 2 focuses on neural models as contextually-aware simulations of language comprehenders. I obtain surprisal estimates of information rate from neural language models and use these to test psycholinguistic theories of utterance production, which postulate speaker monitoring of information rate and, in turn, of comprehension costs. Findings challenge established hypotheses of rational use of the communication channel, especially in dialogic settings—but, overall, they confirm that strategies of utterance production can be described as efficiently containing the comprehension effort of interlocutors.

Part 3 investigates the potential of neural text generators as models of language production. I test whether generators produce language with statistical properties aligned to those of human productions, and then use them to obtain interpretable measures of information rate which are complementary to those used in Part 2. I conclude by collecting insights from the rest of the thesis into a formal framework for artificial simulations of human-like—efficient, communicatively effective, and audience-aware—language production behaviour.

Titles in the ILLC Dissertation Series:

ILLC DS-2018-07: **Julian Schlöder**
Assertion and Rejection

ILLC DS-2018-08: **Srinivasan Arunachalam**
Quantum Algorithms and Learning Theory

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**
Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks

ILLC DS-2018-10: **Chenwei Shi**
Reason to Believe

ILLC DS-2018-11: **Malvin Gattinger**
New Directions in Model Checking Dynamic Epistemic Logic

ILLC DS-2018-12: **Julia Ilin**
Filtration Revisited: Lattices of Stable Non-Classical Logics

ILLC DS-2018-13: **Jeroen Zuiddam**
Algebraic complexity, asymptotic spectra and entanglement polytopes

ILLC DS-2019-01: **Carlos Vaquero**
What Makes A Performer Unique? Idiosyncrasies and commonalities in expressive music performance

ILLC DS-2019-02: **Jort Bergfeld**
Quantum logics for expressing and proving the correctness of quantum programs

ILLC DS-2019-03: **András Gilyén**
Quantum Singular Value Transformation & Its Algorithmic Applications

ILLC DS-2019-04: **Lorenzo Galeotti**
The theory of the generalised real numbers and other topics in logic

ILLC DS-2019-05: **Nadine Theiler**
Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles

ILLC DS-2019-06: **Peter T.S. van der Gulik**
Considerations in Evolutionary Biochemistry

ILLC DS-2019-07: **Frederik Möllerström Lauridsen**
Cuts and Completions: Algebraic aspects of structural proof theory

ILLC DS-2020-01: **Mostafa Dehghani**
Learning with Imperfect Supervision for Language Understanding

ILLC DS-2020-02: **Koen Groenland**
Quantum protocols for few-qubit devices

ILLC DS-2020-03: **Jouke Witteveen**
Parameterized Analysis of Complexity

ILLC DS-2020-04: **Joran van Apeldoorn**
A Quantum View on Convex Optimization

ILLC DS-2020-05: **Tom Bannink**
Quantum and stochastic processes

ILLC DS-2020-06: **Dieuwke Hupkes**
Hierarchy and interpretability in neural models of language processing

ILLC DS-2020-07: **Ana Lucia Vargas Sandoval**
On the Path to the Truth: Logical & Computational Aspects of Learning

ILLC DS-2020-08: **Philip Schulz**
Latent Variable Models for Machine Translation and How to Learn Them

ILLC DS-2020-09: **Jasmijn Bastings**
A Tale of Two Sequences: Interpretable and Linguistically-Informed Deep Learning for Natural Language Processing

ILLC DS-2020-10: **Arnold Kochari**
Perceiving and communicating magnitudes: Behavioral and electrophysiological studies

ILLC DS-2020-11: **Marco Del Tredici**
Linguistic Variation in Online Communities: A Computational Perspective

ILLC DS-2020-12: **Bastiaan van der Weij**
Experienced listeners: Modeling the influence of long-term musical exposure on rhythm perception

ILLC DS-2020-13: **Thom van Gessel**
Questions in Context

ILLC DS-2020-14: **Gianluca Grilletti**
Questions & Quantification: A study of first order inquisitive logic

ILLC DS-2020-15: **Tom Schoonen**
Tales of Similarity and Imagination. A modest epistemology of possibility

ILLC DS-2020-16: **Ilaria Canavotto**
Where Responsibility Takes You: Logics of Agency, Counterfactuals and Norms

ILLC DS-2020-17: **Francesca Zaffora Blando**
Patterns and Probabilities: A Study in Algorithmic Randomness and Computable Learning

ILLC DS-2021-01: **Yfke Dulek**
Delegated and Distributed Quantum Computation

ILLC DS-2021-02: **Elbert J. Booij**
The Things Before Us: On What it Is to Be an Object

ILLC DS-2021-03: **Seyyed Hadi Hashemi**
Modeling Users Interacting with Smart Devices

ILLC DS-2021-04: **Sophie Arnoult**
Adjunction in Hierarchical Phrase-Based Translation

ILLC DS-2021-05: **Cian Guilfoyle Chartier**
A Pragmatic Defense of Logical Pluralism

ILLC DS-2021-06: **Zoi Terzopoulou**
Collective Decisions with Incomplete Individual Opinions

ILLC DS-2021-07: **Anthia Solaki**
Logical Models for Bounded Reasoners

ILLC DS-2021-08: **Michael Sejr Schlichtkrull**
Incorporating Structure into Neural Models for Language Processing

ILLC DS-2021-09: **Taichi Uemura**
Abstract and Concrete Type Theories

ILLC DS-2021-10: **Levin Hornischer**
Dynamical Systems via Domains: Toward a Unified Foundation of Symbolic and Non-symbolic Computation

ILLC DS-2021-11: **Sirin Botan**
Strategyproof Social Choice for Restricted Domains

ILLC DS-2021-12: **Michael Cohen**
Dynamic Introspection

ILLC DS-2021-13: **Dazhu Li**
Formal Threads in the Social Fabric: Studies in the Logical Dynamics of Multi-Agent Interaction

ILLC DS-2022-01: **Anna Bellomo**
Sums, Numbers and Infinity: Collections in Bolzano's Mathematics and Philosophy

ILLC DS-2022-02: **Jan Czajkowski**
Post-Quantum Security of Hash Functions

ILLC DS-2022-03: **Sonia Ramotowska**
Quantifying quantifier representations: Experimental studies, computational modeling, and individual differences

ILLC DS-2022-04: **Ruben Brokkelkamp**
How Close Does It Get?: From Near-Optimal Network Algorithms to Suboptimal Equilibrium Outcomes

ILLC DS-2022-05: **Lwenn Bussière-Carae**
No means No! Speech Acts in Conflict

ILLC DS-2023-01: **Subhasree Patro**
Quantum Fine-Grained Complexity

ILLC DS-2023-02: **Arjan Cornelissen**
Quantum multivariate estimation and span program algorithms

ILLC DS-2023-03: **Robert Paßmann**
Logical Structure of Constructive Set Theories

ILLC DS-2023-04: **Samira Abnar**
Inductive Biases for Learning Natural Language

ILLC DS-2023-05: **Dean McHugh**
Causation and Modality: Models and Meanings

ILLC DS-2023-06: **Jialiang Yan**
Monotonicity in Intensional Contexts: Weakening and: Pragmatic Effects under Modals and Attitudes

ILLC DS-2023-07: **Yiyan Wang**
Collective Agency: From Philosophical and Logical Perspectives

ILLC DS-2023-08: **Lei Li**
Games, Boards and Play: A Logical Perspective

ILLC DS-2023-09: **Simon Rey**
Variations on Participatory Budgeting