UNIVERSITÉ DU QUÉBEC À MONTRÉAL

A THEORETICAL MODEL OF LINGUISTIC PREDICTION: HOW CONTRIBUTIONS FROM VARIOUS LEVELS OF SEMANTIC GRANULARITY INTERACT WITH CONTEXTUAL REPRESENTATIONS WHEN PREDICTING AN UPCOMING WORD

THESIS

PRESENTED

AS PARTIAL REQUIREMENT

TO THE PH.D IN LINGUISTICS

BY

MAXIME CODÈRE CORBEIL

FEBRUARY 2023

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

UN MODÈLE THÉORIQUE DE LA PRÉDICTION LINGUISTIQUE: COMMENT LA CONTRIBUTION PROVENANT DE DIFFÉRENTS NIVEAUX DE GRANULARITÉ SÉMANTIQUES INTERAGIT AVEC LES REPRÉSENTATIONS CONTEXTUELLES LORS DE LA PRÉDICTION DU PROCHAIN MOT

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN LINGUISTIQUE

PAR

MAXIME CODÈRE CORBEIL

FÉVRIER 2023

ACKNOWLEDGEMENTS

This thesis would not have been possible without my thesis supervisor, Elizabeth Allyn Smith, who supported me throughout this process. I would also like to thank my thesis committee members, Giosuè Baggio, Serge Robert, and Grégoire Winterstein, who offered guidance and assistance. Thanks to the members of the linguistics department at UQAM, whom I have worked with throughout this process and who brought me a lot on an academic and human level.

During my stay at University College London, I met several people who helped shape my understanding of pragmatics. Thanks to Richard Breheny, Alison Hall, Paula Rubio Fernandez, Felicity Deamer. I especially thank Robyn Carston for her great insight and thoroughness.

I would also like to thank the granting agencies which allow graduate students to pursue their research while participating in academic life through publications and conferences. Thanks to the SSHRC, the CRBLM, the CRLEC, the NSF, the Faculté des Sciences Humaines at UQAM.

Most importantly, I would like to thank my family, friends, and, more particularly, Valérie and my children, who kept me motivated right until the end. Throughout this doctorate, they have witnessed my academic progress, but also my personal journey.

Finally, I encourage future students to continue seeking and questioning established ideas to advance human knowledge further and promote a collaborative, innovative, and modern science.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

RÉSUMÉ

Dans cette thèse, je développe un nouveau modèle théorique qui prédit le prochain mot d'une phrase. Ce modèle s'inspire de plusieurs disciplines académiques et intègre différents cadres et outils de la linguistique théorique, des sciences cognitives, de la linguistique computationnelle et des modèles du raisonnement analogique. En utilisant une perspective hautement interdisciplinaire concernant la nature de la prédiction linguistique et les types de processus cognitifs qui y sont impliqués, je présente un ensemble de desiderata cognitifs que les théories linguistiques doivent prendre en compte : incrémentalité, non-monotonie et interprétabilité du contenu sous-propositionnel. Je distingue deux types de contributions lors de la dérivation d'une prédiction linguistique : celles provenant de différents niveaux de granularité sémantique et celles provenant de la coordination de l'interaction linguistique et je présente un modèle de langage qui marie ces deux contributions.

Cette approche est testée à la fois pour l'adéquation empirique et pour le réalisme cognitif. Afin de répondre aux contraintes d'adéquation empirique, nous avons vérifié que les prédictions du modèle reflètent les résultats d'études empiriques sur la procédure de cloze. Lors d'une tâche de cloze, un participant se voit présenter une phrase (ou une série de phrases) où des mots ont été omis, et le participant est ensuite invité à compléter le mot manquant. Par exemple, si on montre à un participant une phrase comme "J'ai posté ma lettre, mais j'ai oublié de mettre le...", il est relativement facile de deviner que le prochain mot sera probablement *timbre* et non *voiture*. Une fois que plusieurs participants ont accompli la même tâche de cloze, nous pouvons attribuer une valeur de prédictibilité à chaque mot enregistré en fonction de leur fréquence d'utilisation. La prédictibilité est souvent utilisée en psycholinguistique et en neurolinguistique pour mesurer les propriétés liées à la prédiction et au traitement linguistique; elle a été liée avec le temps de lecture en psycholinguistique et avec la valeur des composants N400 dans les expériences EEG en neurolinguistique. Cette thèse modélise ces valeurs de prédictibilité à l'aide d'outils statistiques et informatiques pour prédire les continuations les plus probables pour une phrase donnée en fonction du sens de cette phrase et, surtout, de la sémantique du discours précédent.

Dans la théorie développée dans cette thèse, les continuations sont calculées à l'aide d'un réseau sémantique basé sur l'activation où le niveau d'activation de tout concept à un moment donné représente le degré auquel ce dernier est activé

par les informations extraites de la phrase tronquée et par le contexte global. Cette valeur d'activation est proportionnelle au poids des connexions entre ces concepts et elle peut être traitée comme une probabilité de cooccurrence entre deux mots. À un instant donné, ces probabilités de cooccurrence déterminent la prédiction linguistique qui est basée sur l'interrelation entre tous les concepts représentés dans le réseau sémantique. Je dérive les réseaux sémantiques à partir des matrices de similarité qui représentent la similarité de cooccurrence entre différents niveaux de constructions linguistiques.

Lors de l'attribution d'une probabilité relative d'occurrence pour les continuations potentielles, nous considérons à la fois la contribution de la phrase tronquée et la contribution du contexte global. J'ai développé des modèles pour deux types d'informations contextuelles : un modèle de *topic* et un modèle de *situation*, et je présente une représentation multicouche de la prédiction linguistique qui intègre la contribution des représentations au niveau de la phrase, la contribution du niveau contextuel et la constante interaction entre eux. Les deux niveaux de représentation ont un rôle primordial dans la dérivation de cette prédiction.

Le modèle de prédiction linguistique présenté dans cette thèse est centré sur la coordination de l'interaction linguistique, et il illustre le lien crucial entre les niveaux de représentation impliqués dans le traitement pragmatique.

Mot-clés: prédiction linguistique, représentation du contexte, influence du contexte, tâche de cloze, espace conceptuel

ABSTRACT

This thesis develops a new theoretically-driven model that predicts the anticipation of upcoming words in a sentence. This model draws from a number of academic disciplines, and it incorporates different frameworks and tools from theoretical linguistics, cognitive science, computational linguistics, and computational models of analogical reasoning.

Using a highly interdisciplinary perspective regarding the nature of linguistic prediction and the kinds of cognitive processes involved therein, I present a set of cognitive desiderata that linguistic theories must consider: incrementality, non-monotonicity, and interpretability of sub-propositional content. I differentiate two kinds of contributions when deriving a linguistic prediction: those coming from different levels of semantic granularity and those coming from the coordination of linguistic interaction, and I present a language model that marries these two contributions.

This approach is tested for both empirical adequacy and cognitive realism. In order to respond to the constraints of empirical adequacy, we verified that the model output mirrors the results of empirical studies on the cloze procedure. A cloze procedure is a task where a participant is presented with a sentence (or a series of sentences) where words have been omitted, and the participant is then asked to complete the missing word. For example, if a participant is shown a sentence like "I posted my letter, but I forgot to put the...", it is relatively easy to guess that the next word will probably be *stamp* and not *car*. When many participants have completed the same cloze task, we can assign a predictability value to every recorded word based on their frequency of use. Predictability is often used in psycholinguistics and neurolinguistics to measure properties related to prediction and linguistic processing, and it has been correlated with processing time in reading time studies in psycholinguistics and also with the N400 components of EEG experiments in neurolinguistics. This thesis models these predictability values using statistical and computational tools to predict the most likely continuations for a given sentence based on the meaning of that sentence and, notably, the semantics of prior discourse.

In the theory developed in this thesis, the possible continuations are obtained using an activation-based semantic network where the level of activation of any

concepts at a particular time represents the degree by which they are triggered by the information retrieved from the truncated sentence and the global context. This relative value of the spreading activation is proportional to the connection weight between these concepts, which can be treated like a probability of co-occurrence between two words. At any given time, these co-occurrence probabilities determine the linguistic prediction based on the relationships between all the concepts represented in the semantic network. I derive the semantic network from similarity matrices representing the similarity of co-occurrence between different linguistic constructions.

When assigning a relative probability of occurrence for potential continuations, we consider both the contribution from the truncated sentence and the contribution from the global context. I developed models for two kinds of contextual information: a topic model and a situation model, and I present a multi-layered representation of linguistic prediction that integrates the contribution from the sentence-level representations, the contribution from the contextual level, and the constant interaction between them. Both representational levels have a primordial role in the derivation of this prediction.

The model of linguistic prediction presented in this thesis is centered around the coordination aspect of linguistic interaction, and it illustrates the crucial connection between the representational levels involved in pragmatic processing.

Keywords: linguistic prediction, representation of the context, contextual influences, cloze task, conceptual space

INTRODUCTION

This resolutely integrative thesis investigates linguistic anticipation, namely how we anticipate the next word of a sentence, bringing together different notions and approaches to propose a theoretical processing model for linguistic prediction.[1]

First, we consider the cloze task. The most common version of a cloze task is a task in which participants are asked to complete a sentence that has been truncated. For example, participants are shown sentences like "I posted my letter, but I forgot to put the...". In this particular sentence, it is relatively easy to guess that the next word will most probably be *stamp* and not *car* or something else. When many participants have completed the same cloze Task, it is possible to assign a cloze score (or cloze value) to every recorded word based on their frequency of use. Once we have computed the different cloze scores for different incomplete sentences, we can use this information in different experimental settings to measure different properties related to linguistic prediction and sentence processing.

Secondly, *predictability* has been defined as the easiness one can predict the upcoming word given a specific linguistic context. Predictability is often used in both psycholinguistic and neurolinguistic empirical studies. It has been correlated with processing time in reading time studies in psycholinguistics, and it has been correlated with the N400 components of EEG experiments in neurolinguistics. One problem is that the notion of predictability seems to overlap many other notions

[1]The theoretical framework I am presenting in this thesis should be understood as a general theory of prediction that could be used to develop different concrete models of linguistic prediction. Throughout this thesis, I use the word *model* to refer to the more specific implementation of this theoretical framework I am presenting here.

like those of plausibility and possibility. Thus, a redefinition of the concept of predictability or at least a full investigation of the exact nature of the cloze task itself would be beneficial.

Thirdly, computational approaches of meaning and many tools have been recently developed to represent meanings of words and sentences at different detail levels. For example, Distributional Semantic and neuronal network approaches are now rapidly improving, and some recent cognitive models of surprisal (or prediction) are interested in simulating empirical results. The recent advances in these three fields make it possible to combine their respective contributions to develop an integrative view of linguistic processing.

For this purpose, I first present the basics regarding the nature of the cloze task and the kinds of cognitive processes involved (Chapter 1). I then discuss the current state of empirical research about prediction and processing time regarding the role of predictability and cloze scores (Chapter 2). Chapter 3 presents a desideratum for linguistic tools, and I present different linguistic approaches that fit well with these requirements. Chapter 4 presents a language model that integrates the contribution from meaning compositions. The contribution from the coordination aspect of linguistic prediction and its relationship with the meaning compositions is discussed in Chapter 5. Finally, in Chapter 6, I discuss issues and potential improvements for the model, and I compare its structure and characteristics with other related approaches interested in linguistic prediction.

CHAPTER I

WHAT IS LINGUISTIC ANTICIPATION?

We can start by asking ourselves one question: is it possible for someone to predict or anticipate upcoming words during day-to-day communication? One might say no, because we normally do not know in advance what our interlocutor will say. It is indeed difficult to precisely predict what one wants to utter without any other clues. However, surprisingly enough, it is not uncommon to do so, at least at the utterance level. Consider, for example, (1) where the last word of the utterance is missing:

(1) I went to the bakery for a loaf of ...

When we ask different people to find the missing word of (1), they almost all say that this word should be *bread*. This task has been completed by 400 participants, and 98% of them chose *bread* as the missing word (Block & Baldwin, 2010). This result shows that even if we usually think that we cannot predict the next upcoming word, we can still do so, and it seems that our prediction is very similar to other's predictions.

Using examples like (1) we could even question the optionality of this prediction. If I utter (1) in a conversation and I stopped right before the last word, you might

already have predicted the next word, and you would have difficulties not predicting it. In other words, the hearer cannot not predict the next upcoming word when facing incomplete utterance like (1). Another property of this prediction is the uniformity of responses made by different participants, i.e., 98%.

The sentence in (1) is called a High-Constraining Utterance (HCU) because it triggers specific predictions about upcoming words (Grisoni et al., 2017; Kuperberg et al., 2020): the different information that was expressed by (1) made it possible to narrow down the possibilities to a single word. During this predictive process, we might start our search for the next word using the fact that not all kinds of lexical categories can follow a preposition like *of*. Usually, in English, after a preposition, we have an adjective or a noun phrase that already constrains the category of the upcoming word because it tells us that we should not look for a verb, for example. Then, we can use the fact that the lexical meaning of the word *loaf* is almost always associated with *bread*, so if we want to continue the constituent 'a loaf of...' we will most probably use *bread* instead of something else like *meat*. To definitively reject *meat* as the continuation, we could use the lexical meaning of the word *bakery*, which, once associated with *loaf*, will surely be enough to select *bread* as the expected continuation.[2]

Unfortunately, highly-constraining utterances like (1), where almost everyone predicts the same word, are relatively rare compared with less constraining ones like (2).

(2) The kind old man asked us to ...

---

[2]Throughout this thesis, the term *continuation* refers to the word that follows the point of truncation of a given sentence. It should not be mistaken with the concept of "continuations" used in theoretical computer science that refers to the context surrounding an expression (Barker & Shan, 2014).

Here, it is more difficult to predict the missing word because we have more options that would be regarded as an acceptable continuation of this utterance. To illustrate this difficulty, we can try to reproduce the same predictive process I described for (1). First, we note that the kind of word that follows *to* is usually a verb or a noun phrase, but here if we consider the bigger constituent 'asked us to' instead, then the next word is definitely going to be a verb. The problem here is that we cannot use the meaning of any word in the rest of the utterance to help constrain the domain of possibilities because *kind*, *old*, *man*, *asked* are not strictly associated with one single word. In this case, it would help us consider the larger context of the utterance, but since we only have access to (2), we remain in an uncertain world, having to guess more than predict the next upcoming word. This uncertainty is reflected by the many responses gathered by Bloom & Fischler (1980). Here the continuations were more diverse: *stay* (26%), *help* (21%), *leave* (10%), *dinner* (5%). An utterance that is not constraining enough to the next word is called a Low Constraining Utterance (LCU).

The distinction between the two cases (LCU versus HCU) is related to the number of clues available to help derive our prediction. In this thesis, I am interested in describing how these clues are combined to influence our predictions.

Before presenting evidence for the presence of anticipatory behavior during linguistic processing, it is important to distinguish between anticipation and prediction. As defined by Van Petten & Luka (2012), 'prediction' is used when the comprehender anticipates a specific word (or lexical item) from the sentence, and 'anticipation' is used as a broader umbrella term to indicate that a reader/listener anticipates some semantic content, and may or may not have narrowed that expectation to a particular word (Van Petten & Luka, 2012). Although worded differently, this distinction is close to the one proposed by Lederer (1978). Lederer described two types of anticipation: linguistic anticipation and sense expectation

(or extra-linguistic anticipation). The linguistic anticipation is based on language knowledge and is used to predict the end of sentences like (3-a). Whereas for the sense expectation, we need extra-linguistic information because it cannot be completed from linguistic cues only as we seen in (3-b) (Vandepitte, 2001).

(3) a. She was green with...
b. They held off...

There is certainly a parallel to make between Lederer's distinction between linguistic anticipation and sense expectation and what was described in the first section as High-constraining and Low-constraining utterances (HCU and LCU). As was the case for the HCU, we would generally be able to predict that the next word of (3-a) would most probably be *envy*, but we would have difficulty predicting the continuation of (3-b) as it was the case for the LCU case because we lack clues that would constrain the possibilities. In the latter case, we need extra-linguistic information from the larger context to derive this prediction.

What has been divided procedurally as anticipation based on semantics and anticipation based on pragmatics by Lederer is instead viewed as a difference in the kind of results, either a specific word or a vaguer idea of semantic content. In this thesis, I use *prediction* when referring to a specific lexical prediction and *anticipation* when considering a broader, more general view of prediction.

1.0.1 Anticipation and linguistic processing

Le Ny (1978) suggested that anticipation was an ongoing cognitive activity run-

ning in parallel to the perception of the incoming speech sounds and their semantic analysis; Kalina (1992) argued that anticipation was part of processing a source text. In a famous series of experiments, Tanenhaus et al. (1995) measured the gaze of interlocutors when spoken to. Their results showed that the interlocutor established a visual reference to an object during the processing of linguistic data, i.e., the interlocutor anticipated the referents of some word even before this word had been fully pronounced. In their experiments, they asked participants to move different objects around by using utterances like (4).

(4) Pick up the candy. Now put it above the fork.

For this particular case, Tanenhaus et al. (1995) showed the participants' eye gaze was initiated towards the candy approximately 145 ms after the end of the pronunciation of the word *candy*. Because it takes about 200 ms to change the direction of the eye gaze, this means the identification of the object 'candy' precedes the complete pronunciation of the word *candy*. In another example, they asked a participant to move a card as in (5).

(5) Put the five of hearts that is below the eight of clubs above the three of diamonds.

Here, the participant looked at an eight of clubs that was above the five of hearts immediately after *below the*. This particular example showed that a person does not wait until the whole meaning of a sentence has been uttered before processing the information available at any given moment. This evidence points towards an incremental and greedy process of the interpretation of linguistic meaning. Therefore, a person is not only able to integrate incomplete information, but she

does it at the earliest moment possible, as soon as she can establish a reference even if it has not been explicitly stated yet (Tanenhaus et al., 1995). Put together, the results for (4) and (5) support the idea that a person anticipates the next upcoming bit of information.

This human propensity for anticipation was also used to explain evidence from Sedivy et al. (1999) which showed that adjectives were also processed incrementally. Their experiment asked participants to grab different objects among four different objects (one pink comb, one yellow comb, one yellow bowl, and one knife) while registering their gaze.

(6) Touch the pink comb. Touch the yellow ... (comb/bowl).

Their results indicated that most of the time, the modified noun was interpreted contrastively, i.e., participants were looking at the comb and not the bowl, and this implied that the adjective *yellow* was processed as soon as uttered. Their results provide additional evidence that linguistic processing is incremental and that interpretation is derived moment-by-moment (Sedivy et al., 1999).

Evidence for incremental processing has also been observed beyond the sentence level. In Rohde & Horton (2014), they were interested in coherence relations between utterances, and their results disconfirmed the Clausal Integration account, which states the "intersentential pragmatic relationships can only be made after the structural and semantic properties of the two individual sentences have been determined" (Rohde & Horton, 2014, p.669). Instead, they showed that anticipatory looks revealed a preference for particular coherence relations, such as *Cause* and *Consequence* relations, even before the end of the utterance. For this purpose, they trained participants to look at different locations, each representing different

coherence relations. The participant's gaze was measured when they read sentence continuations that violate the verb-based expectation for coherence relations.

(7) Prompt: Arthur scolded Patricia in the hallway. [Cause bias]

a. Cause continuation: She had put thumbtacks on the teacher's chair.

b. Occasion continuation: He then sent her to the principal's office.

(8) Prompt: Heidi shipped Eric a package. [Occasion bias]

a. Occasion continuation: He wrote her a thank you note.

b. Cause continuation: She thought he'd like some cookies from home.

They showed that the looking behavior before the verb and the looking behavior after the verb were not the same, and this change was sensitive to verb classes. These results demonstrated that establishing a coherence relation is an expectation-driven process because participants were trying to predict the coherence relations from the available cues from the first sentence before having processed the second sentence.

Finally, we could also mention the event-related brain potential (ERP) experiment by Van Berkum et al. (2005) where they observed an effect coming from the gender of the preceding article of an expected word. ERP experiments will be introduced in more detail in Chapter 2, but we can still briefly look at their results here. The participants were presented Dutch stories that supported a specific noun's prediction, but these stories were continued with a gender-marked adjective whose suffix mismatched the upcoming noun's syntactic gender (Van Berkum et al., 2005).

(9) The burglar had no trouble locating the secret family safe.

a. Of course, it was situated behind a big_{Neuter} but unobtrusive painting. (consistent)

b. Of course, it was situated behind a big_{Common} but unobtrusive bookcase. (inconsistent)

Van Berkum et al. (2005) observed that ERP effects were larger when the adjective was inconsistent with the expected upcoming word and that reading time was slower for these cases. Their results suggest that a prediction about the upcoming word is already derived when the participant is processing the adjective, and this implies that prediction also play an active role during the incremental linguistic process.[3]

Put together, this evidence from Tanenhaus et al. (1995), Sedivy et al. (1999), Rohde & Horton (2014), and Van Berkum et al. (2005) support an incremental view of linguistic processing, and they are also compatible with the idea that the role of anticipation is essential for linguistic processing.[4] The fact that linguistic processing is incremental allows the hearer to process every new input as soon as possible, which would imply that if an input is processed faster, it should be beneficial for the hearer. Anticipation would then be a natural way to shorten the processing time because an input that is correctly predicted is an input that is already partly processed.[5]

---

[3]If we are not presenting a non-incremental point of view it is because there is no dispute regarding the incrementality of sentence processing, only different ways of modelling the granularity of this incrementality.

[4]Observations about participants being able to anticipate Discourse Relations (Scholman et al., 2017) are also pointing in the same direction.

[5]As we will discuss in Chapter 5 and Chapter 6, predictions are derived in parallel at different spatial and temporal scales which means that even though a prediction at a particular scale turns out wrong, the error signals that is propagated across scales will eventually be suppressed by a higher-level prediction (Clark, 2016; Friston, 2005). In other words, the cost

As argued by Kamide et al. (2003), incremental processing and anticipation are independent of one another, and it is essential to mention that anticipatory behavior does not diminish the importance of the interpretation process in itself because, in the end, an input has still to be interpreted, and it is the output of this interpretation process that is then compared with the information that the system has predicted. In the words of Van Berkum et al. (2005, p.464): "predicting the trajectory of a frisbee does not preclude actually catching it."

Furthermore, anticipation could also naturally explain current results in turn-taking studies (De Ruiter et al., 2006; Levinson & Torreira, 2015; Levinson, 2016; Corps et al., 2018). In a similar fashion to what Tanenhaus et al. (1995) proposed, Levinson & Torreira (2015) argued that anticipation is a key explanation when trying to make sense of the short gap between speaking turns. Namely, the gaps between speaking turns are very short (around 200 ms), but the latencies involved in language production are much longer (600 ms). Positing the existence of some anticipatory behaviors related to language processing would easily explain this discrepancy between short gaps and longer latencies because if the second speaker can anticipate the complete utterance of the first speaker, then she is also able to plan her response (Holler et al., 2015; Levinson, 2016; Riest et al., 2015). In other words, a speaker starts thinking about the next utterance even before the previous speaker has finished his utterance.

To better understand the mechanics of the anticipation process during turn-taking, Corps et al. (2019) used constrained and unconstrained utterances during a dialogic simulation (see (10-a) and (10-b)) to show that linguistic prediction did not influence the anticipation of the end of speaking turn. Although, it seems that the broader discourse context plays a more critical role when a listener tries to

---

of recovering a wrongful prediction will be counterbalance by the benefits of having a correct prediction at a higher level.

anticipate the end of a speaking turn: they found that listeners responded earlier in the constrained case ((10-b)) compared to the unconstrained case ((10-b)).

(10) a. [Constrained]: What music do you listen to? I really like Taylor Swift.

b. [Unconstrained]: Is there anything specific I should know about you? I really like Taylor Swift.

These results support the idea that anticipation involves processes that are taking place above the sentence level.

1.0.2 Anticipation or Prediction?

Anticipation has also been discussed in other fields related to communication in general, e.g., in the domain simultaneous interpretation (SI). In SI, the interpreter, who is translating from one language to another on the fly, has to form a complete proposition from a truncated one, i.e., from a proposition that has yet to be expressed entirely (Le Ny, 1978; Lederer, 1978).

(11) sugars are playing important positive roles.
Le role positif et important des sucres.

In this example taken from Vandepitte (2001), an expert interpreter can produce the word *role* in French even before the word *roles* has been pronounced in English. This need for anticipatory SI behavior has already been studied by Chernov (2004) and also by Vandepitte (2001). Vandepitte sees anticipation as producing

a particular mental representation without being completely induced yet through the speaker's sounds (Vandepitte, 2001, p.38).

This kind of anticipatory behavior is very close to what was described by Lau et al. (Lau et al., 2013, p.487) as the "generation of expectancies from contextual representations held in working memory." If the role of anticipation in SI is well accepted, its role in a 'standard-setting,' where two individuals talk to each other, is still debated. For example, in Vandepitte's view (2001), anticipation involved in SI should not be mistaken with a "normal comprehension process" because, in the latter, the hearer does not have to engage in anticipation. In contrast, the interpreter always has to anticipate to be better at his task. The difference between SI and 'normal' conversation is the difference in the goal of the anticipatory process.

An interpreter's goal is not to translate one-to-one the meaning of every word but to convey the idea behind the sentence. The interpreter has to make sure she translates as fast as possible the uttered content to minimize the lagging between the source utterance and the translated one. On the other hand, in a 'normal' conversation, the hearer's goal is generally to retrieve the meaning conveyed by the speaker, and this might involve being able to vaguely anticipate what a speaker says during his speaking turn, and, in that case, the temporal constraint is less important (Ferreira & Lowder, 2016).

The idea defended by Vandepitte (2001) that anticipation is motivationally driven or goal-oriented is in phase with the recent empirical results from Brothers et al. (2017) where they measured different brain responses by varying the instructions to the participants. In the first part of their experiment, they asked the participant to read a text passage, while in the second part, they asked them to predict each passage's final word. Brothers et al. (2017) showed the different brain activation

patterns relative to these two instructions, which in turn implies that the nature of the goal of a task influences the way anticipatory behavior is used.

Notwithstanding this significant result, the question here should not be about the influence of different goals on anticipation, but rather it should be about the possibility that anticipation is a strategy that could be applied at the comprehension stage (Kalina, 1992). The most common reason to justify the need to have anticipatory behavior is that it should help minimize the cognitive effort to process upcoming information because predicting the upcoming information should facilitate the integration of this information during language comprehension (Ferreira & Chantavarin, 2018).

In recent years, this question about the role of anticipation in language processing has drawn much attention from the linguistic and cognitive science communities (Kuperberg & Jaeger, 2016), but despite being increasingly investigated, empirical measures for the role of such predictive features (Jaeger & Snider, 2013; Lau et al., 2013) were inconclusive, and, as of now, the exact role of predictive behavior in linguistic processing is still debated (Ferreira & Chantavarin, 2018; Huettig & Mani, 2016; Nieuwland, 2019).[6]

1.1 Cloze task

Having introduced linguistic prediction and some empirical data that motivate it, we turn to this thesis's main subject, namely the cloze task. In linguistics, the interest in the anticipation of an upcoming word has grown recently, but

[6]Huettig (2015), for example, argues that even if anticipation is an essential aspect of language processing, it may not be one of its fundamental principles.

almost all data sources on the subject are based on the classic article by Taylor (1953). In this paper, Taylor introduces the 'cloze procedure' to measure the cloze probability of a word. A cloze procedure is a task where a participant is presented a sentence (or a series of sentences) where words have been omitted, and the participant is asked to complete the missing word (Bloom & Fischler, 1980). All the responses for every incomplete sentence are then weighted according to their respective frequency of occurrence, and this proportion is called the cloze probability (or predictability) of a word given this particular context.

Over the years, this kind of cloze task was very often used, and many variations of the instructions have been observed: some participants are asked to find the most natural continuation, the most plausible, the 'best' one, or they are asked to name the word that first come to mind when wanting to complete a truncated sentence (Staub et al., 2015, p.2). In all these cases, even if the instruction might vary slightly, the gist of the experimentation is the same: to retrieve the most predictable word given a preceding context.

(12) He loosened the tie around his ...

For example, the sentence in (12) was completed by *neck* 96% of the time in the Bloom & Fischler (1980) study and 97% of the time in the Block & Baldwin (2010) study. In this linguistic context, the word *neck* has a very high cloze probability. Generally, a high cloze score is attributed to results that occurred more than 66 percent of the time, while the words below 66 percent are described as low cloze scores (Block & Baldwin, 2010). Predictability is assessed from the cloze probability value by comparing different words for the same incomplete sentence and creating a predictability scale accordingly.

Another way to measure the effect of predictability is to compare the reading time of the same word for two different contexts, one in which the word has a high cloze value and the other one for which the cloze value is low. It is now widely accepted that predictability is correlated with reading time, which means that a word with lower predictability should take longer to process (Staub et al., 2015).[7] Even if this strong correlation has already been argued for over the past years using different results that measured different aspects of expectation-driven processes: lexical, semantic, syntactic, pragmatic, we still do not know much about what is achieved by the participant during a cloze task (Smith & Levy, 2013).

1.1.1 Cloze tasks in linguistics

Apart from measuring cloze probability values, cloze tasks are often used in linguistics to investigate different linguistic phenomena such as pronoun production and anticipation of thematic roles. One illustration of these completion tasks is given in Kaiser & Cherqaoui (2016):

(13) a. Aurélie a bousculé Thérèse hier au cinema, alors... celle-ci s'est mise à pleurer.
'Aurélie shoved Thérèse yesterday at the movies, so... this one started to cry.'

b. Arnaud a battu Pascal pendant la soirée chez des amis, et après ... il lui a présenté ses excuses.
'Arnaud beat Pascal during the evening with friends, and then ... he

[7] I will come back to this notion of reading time in Chapter 2.

apologized to him.'

c. Philippe a poussé Jacques dans l'escalier Dimanche, et après ... celui-ci s'est fait mal en tombant.
'Phillippe pushed Jacques on the stairs on Sunday, and then ... this one hurt himself while falling.'

In their experiment, Kaiser & Cherqaoui (2016) manipulated the form of the anaphoric expression (*she*, *this one*) and the connective between the two clauses (*then*, *as a result*). They showed that pronouns tend to be "interpreted as referring to subjects, and anaphoric demonstratives tend to be interpreted as referring to objects" (Kaiser & Cherqaoui, 2016, p.66).[8] This result is fascinating because it supports the idea that certain words are more likely to be associated with certain kinds of constituents in a sentence, without any regard to the actual sentence. In other words, it shows that participants seem to have an anticipative bias towards the use of pronouns and the use of demonstratives concerning what they could refer to.

If initial models of pronoun production were centered around expectation-driven processes and coherence relations (Kehler, 2002), more recent models are deliberately focussing on finding a way to understand how these two components, namely the comprehender's expectations and the production-based anticipatory bias, play a role in pronoun interpretation (Kaiser, 2013).[9]

Anticipation about longer dependencies has also been observed, i.e., anticipation about goals and thematic roles rather than only locally licensed information like

---

[8]This same result was also observed in German (Kaiser, 2011).

[9]More recent models are adopting a Bayesian approach that can link production with interpretation directly (Kaiser & Cherqaoui, 2016; Kehler & Rohde, 2018; Kaiser, 2013; Rohde & Kehler, 2014; Kehler et al., 2008).

an object or a subject (Kamide et al., 2003). In this study, Kamide et al. (2003) used 3-place verbs such as *spread* and *slide* where both verbs subcategorized 2 arguments that are located after the verb itself, the third argument being a goal: *Spread(Agent,Theme,Goal).*

(14) a. The woman will spread the butter on the bread.
b. The woman will slide the butter to the man.

By measuring eye gaze, they showed that participants were able to anticipate the goal of the action in advance as they looked more at the man when they encountered sentences like (14-b) (Kamide et al., 2003). Once again, it seems that anticipation is possible even just after the verb, which is another evidence for an incremental interpretative process.

Finally, these results are also in line with those of Rosa & Arnold (2017) where they found that pronoun production was more prominent when referring to a goal than when referring to a source. In their experiment, they contrasted causal statements like (15) and transfer-of-possession events like (16), and in both cases, participants interpreted the pronoun to be referring to the implicit cause or the goal of the event (Rosa & Arnold, 2017).

(15) a. The butler blamed the chauffeur because he.... (murdered someone).
b. The butler impressed the chauffeur because he... (figured out the case).

(16) a. The butler gave the threatening note to the chauffeur and he... (turned it in to the police).
b. The butler received a ticking bomb from the chauffeur and he ...

(chucked it into the river).

According to their results, the goal argument is considered more predictable because the chances are, it will be mentioned in the sentence. This result seems like an anticipatory bias towards interpreting a pronoun as primarily referring to a goal instead of a source (Rosa & Arnold, 2017). These results can be easily linked with the one from Kaiser & Cherqaoui (2016) since object interpretation is more associated with causal cues or goals, whereas subject interpretations are more associated with non-causal or temporal uses.[10]

1.1.2 Limitations of the cloze task

The cloze task and the predictability value have been used extensively in the literature in psycholinguistics and neurolinguistics, but that does not mean that we have a detailed understanding of all its intricacies. We can ask a participant to complete a truncated sentence, and then we can measure the predictability of a given word, but, as stated by Smith & Levy (2011), we still know almost nothing about what the nature of this task is. We know that performing a cloze task does not automatically involve the exact mechanisms used for language processing. In other words, asking someone to complete a sentence is not the same as processing an already completed sentence. Completing a cloze task is an offline task where the participant has time to think and decide the best response possible, which is not the case for standard language processing (Smith & Levy, 2011). Predictability

[10] In Chapter 3, this anticipatory bias will be translated into a correspondence function between a speaker (or an interpreter) and a conveyed meaning, and its influence will further be discussed in Chapter 5.

is thus an offline measure, and this has to be taken into account when using the cloze task results in a study about linguistic processing (Chow et al., 2016b). In both cases, it is a matter of using the correspondence between a linguistic representation and a state of the world, but the directionality is not the same.

The distinction between interpretation and prediction will be discussed in Chapter 3, but the critical thing to note here is that it is not clear from the cloze task results alone how the participant makes this prediction or this choice of words. Do they generate their response by probability assessment of the most probable word in the context or by using associative pre-activation (Smith & Levy, 2013)?

If the choice is activation-based, then a cloze task is to be understood as a race to retrieve the first word coming to mind when interpreting the incomplete sentence, i.e., the first word that is activated above some threshold level of activation (Staub et al., 2015). If this choice is expectation-based, i.e., based on the probability of occurrence given the context, then a cloze task is a measure of this conditional probability; it is a sampling from a subjective probability distribution. We thus have two different kinds of probabilities: conditional probability of word occurrence and the probability that a word is reaching the activation threshold first (Staub et al., 2015) and we have no easy way to know which of them is measured during a cloze task.

Another problem with the cloze task is that once a cloze probability (or predictability), has been measured, there are two ways we can interpret the results (Staub et al., 2015). To illustrate the difference between the two, let us suppose two words were used to continue a truncated sentence as shown in (17).

(17) Truncated sentence ...

a. the word *car* has a cloze score of 90%.

b. the word *boat* as a cloze score of 10 %.

The most common interpretation for these results would be that the expectation to encounter the word *car* in the given context are higher than the one for the word *boat*. This higher expectation should be treated as applying to any individual performing a similar task. Therefore, the cloze result here is considered a measure of predictability linking a word with a linguistic context without considering the participants' idiosyncrasies. For example, when participants are asked to provide multiple answers and sort them by probability like in Roland et al. (2012), then these classifications can be used to compute the average preferred response, and one can conclude that this answer generally has the highest predictability among all the participants. Under this interpretation, *car* is more predictable even for participants that did not classify it in the first position in their list (Staub et al., 2015).

The other interpretation for the cloze scores is about individual variability. In this view, the fact that *car* has a cloze score of 90% means that for 90 percent of the participants, it was the best candidate for the continuation of the sentence. Here the variability of cloze scores is due to the variability in personal experience and linguistic knowledge between participants (Staub et al., 2015). Under this interpretation, every participant has only one most expected word, and the cloze score is a measure of the individual variability of the possible response. For example, there is evidence that participants prefer words that are more familiar, shorter, and less formal since they want them to be felicitous in the formal context of an experiment (Smith & Levy, 2011). This discrepancy between these two interpretations of a cloze score is challenging because most data gathering involves multiple trials and averaging the results. In principle, cloze values are compatible with both interpretations (Van Petten & Luka, 2012). In this thesis,

I only consider the first interpretation because it is the most common one, but also because building a personal cloze task simulator is a much bigger and more challenging problem to tackle.[11]

1.1.3 Simulating a cloze task

By choosing the first view, we could treat predictability results as conventionalized cloze scores across a population, making it easier to bring consistency to the model. In this thesis, I am interested in describing a theoretically driven model to explain cloze task results, i.e., to simulate linguistic prediction directly from a truncated sentence. Cloze values are usually measured by asking participants to complete a given sentence and then averaging all the answers, but the goal here is to compute them without outsourcing. Cloze tasks are generally tailored to fulfill specific experimental requirements, and every empirical study has its measures and particular ways of designing a cloze task. In this thesis, I chose as a starting point the well-known study by Bloom & Fischler (1980) where they give completion norms for 329 sentences. A follow-up study by Block & Baldwin (2010) added 398 new sentence completion norms and made their results available for the scientific community's benefit. In the following chapters, I discuss how to develop a model that would simulate linguistic prediction at the word level while also being compatible with the most recent experimental results about linguistic prediction and its effect on linguistic processing.

---

[11] However, this does not mean it is not be feasible to do so. See Chapter 6 for a discussion about this issue.

1.2 Interdisciplinary Nature of this Thesis

We can define the nature of this thesis in terms of Marr's three levels of analysis (Marr, 1982). According to Marr's book, any information-processing task, or any complex system, could be divided into three levels of analysis: computational, algorithmic, and implementational.

- Computational level
  The goal is to describe the strategy or the nature of the process: What does the system do? What is the logic behind this strategy?
- Algorithmic level or representational level
  The goal is to specify the system's algorithm to process the information: How does the system work? What kind of representations does it use?
- Implementational level
  The goal is to understand the physical realization of the algorithmic level: What neural structures or neuronal activities are involved in the process?

The three levels are intertwined, and every level is thought to be a realization of the level before it. To achieve a complete understanding of a system, we thus have to work towards developing integrative models that could encompass all three levels. The goal here is to cover the necessary avenues related to the nature of linguistic prediction in order to develop a tentative model that could fulfill the requirements of an integrated theory.

This thesis is inherently multidisciplinary because predictive behaviors are also inherently multidisciplinary. Language comprehension is often represented at the center of a triangle comprising computational linguistics, information theory, and

cognitive neuroscience because they must all be involved when trying to understand language (Armeni et al., 2017). In this thesis, I use a slightly different schematic and terminological view where linguistic prediction could be modeled with the collaboration of cognitive science, computational approaches, and linguistics, as in Figure 1.1 (Armeni et al., 2017, adapted from their Figure 1).

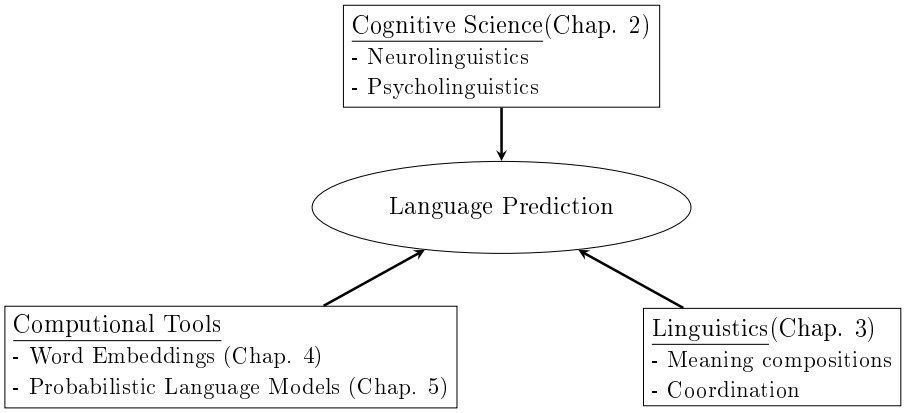


Figure 1.1 A schematic depiction of the interdisciplinary collaboration required to model linguistic prediction

Following this view, to develop a model capable of simulating linguistic prediction, we need to consider all three domains of knowledge because the said model has to respond the constraints from all of these fields. Interestingly, it is possible to establish a relationship between each contribution from these three domains and Marr's three levels. Within this schematic representation, the cognitive science domain provides the empirical data the model should be based on, but it also gives clues about the physical realization of linguistic processing, i.e., the implementational level. The contribution from linguistics resides more at the computational level, where the theoretical constraints of the linguistic processes are determined in a principled way by looking at different linguistic behaviors in communication.

Finally, the tools developed by computational approaches, despite their names, are

closer to the algorithmic level than the computational levels since they provide representations one can work with to develop a model about linguistic prediction. The correspondence between Marr's three levels and the contribution from these three domains of research is far from perfect, but at least it underlines their respective importance, and it provides an opportunity to work towards an integrative theory of language processing/prediction. In the following subsections, I discuss the respective contribution of these three fields, and I present the general organization of this thesis.

1.2.1 Cognitive Science

For this thesis's scope, I consider that all empirical studies interested in how language is cognitively processed are part of cognitive science. In the schematic view that inspired Figure 1.1, Armeni et al. (2017) used 'cognitive neuroscience' instead of 'cognitive science,' and their views would more readily correspond to the implementational level. However, I chose to stick with the more general appellation 'cognitive science' because I think it encompasses more naturally the contribution from neurolinguistics studies and psycholinguistic studies. It could be argued that psycholinguistic results would not be characterized as contributing to understanding the implementational level. However, if we think of the physical constraints for realizing a process, psycholinguistics and neurolinguistics both provide clues to understand better how linguistic prediction works (Hasson et al., 2018), and both offer results upon which implementation models would be evaluated. Chapter 2 is entirely dedicated to discussing empirical results about linguistic processing and linguistic prediction, and these results will serve as a cognitive thread with which I intend to tie the linguistic and computational approaches.

1.2.2 Linguistics

Linguistics is generally described as being divided into different sub-fields such as phonetics, syntax, or semantics. This way of separating sub-fields in linguistics is relevant when one is interested in distinguishing different linguistic units. However, since my purpose is to discuss linguistic processing and how different kinds of information are taken into account during an anticipatory process, I chose to use a different distinction based on the nature of the information used to derive the prediction. Hence, I divide the relationship between linguistic prediction and linguistics into two types of contributions: the linguistic information emerging from the combination of different words or linguistic units, i.e., meaning compositions, and the information derived from the fact that communication, and therefore prediction, are about linguistic coordination between two agents, i.e., coordination processes. This distinction in terms of different contributions to meaning blurs the conventional separation between semantics and pragmatics. The difference between the contribution coming from the meaning compositions and the coordination aspect of linguistic interactions will be presented in Chapter 3, and the intricacies of the interaction between the two will be discussed in Chapter 5. In this thesis, I do not discuss the lower granularity levels of linguistics because I am primarily interested in simulating a prediction at the level of the word and not beyond that. However, this is not to say that phonology could not help form a prediction at the word level since. As argued by Van Berkum & Nieuwland (2019), different levels of granularity all provide their respective contexts, and these contexts all contribute to the understanding of the entire discourse. The approach presented in this thesis is centered around a prediction about an upcoming word, and I do not discuss discriminating units below the level of the word.

1.2.3 Computational Approaches

The third important domain for understanding linguistic prediction is the whole spectrum of knowledge related to the development of new computational approaches to cognition and language. Computational tools (or approaches) pervade both linguistics and cognitive science, and, in this thesis, I focus on two of them: word embeddings and probabilistic models of cognition.

The idea behind word embeddings emerged from the distributional hypothesis (Harris, 1954) that words expressed in similar contexts must have similar meanings. Word embeddings are generally derived indirectly by examining how words and their neighbors are distributed within a corpus of data. Chapter 4 discusses these word embeddings and explains how they can be used to represent the contextual meaning of a word (Gastaldi, 2020). The fact that they are well-defined mathematical objects, namely vectors, makes them helpful when representing meaning because they can be easily combined to form more complex units of meaning.

Finally, probabilistic models, and Bayesian approaches especially, are more popular than ever for explaining empirical data both in psycholinguistics (Jurafsky, 2002; Heller et al., 2016; Kehler & Rohde, 2018) and in neurolinguistics (Nicenboim et al., 2020; Delaney-Busch et al., 2019). Chapter 5 introduces them and explains how they could contribute when taking into account the coordination between two agents.

1.2.4 Summary

This thesis is an integrative effort to benefit from the recent developments within three fields of research: linguistics, cognitive science, and computational approaches. In the following chapters, I look more closely at each of these disciplines to circumscribe their contribution to linguistic prediction and their respective constraints when developing an integrated model of linguistic prediction. This thesis's primary purpose is to develop an explanatory model of linguistic prediction that could help bridge the gap between the higher-computational and the lower-implementational levels.[12] This thesis is in line with the recent tendency to explain how representations are combined within one's mind while using neurological data to develop mechanistically realistic models (Baggio, 2018; Hagoort, 2020; Nieuwland & Martin, 2017).

Additionally, this thesis helps improve our understanding of linguistic processing, which is sometime divorced from computational models, especially in machine learning and deep learning (Linzen, 2019). In a world where machine learning and computational models of prediction have become more and more efficient, it is still of utmost importance that a model of linguistic prediction could help us better understand the cognitive and the linguistic ramification of such process, in addition to being able to perform well in a cloze task. For example, you could turn to a well-known search engine and input any truncated sentence, and it would probably give you a sensible answer, but it would not help you understand how a human mind would have provided you with this answer. In other words, even though computational models based on neural networks are becoming more and

[12]This problem is often called the mapping problem (Poeppel, 2012) because it involves the mapping of representations between language science and cognitive science.

more sophisticated, they are not very informative about what is going on in the mind of a person that has to perform a cloze task. The opacity of neural networks is generally described as a significant limitation when building a cognitively sound model of a given process (Marcus, 2018). Even though some language models have now achieved a high-performance level, it does not mean that they have acquired linguistic knowledge like compositionality (Linzen, 2019). It is the role of linguistics to provide detailed requirements about the capacities necessary for language processing and, in this case, linguistic prediction.

Finally, methods using deep learning have great difficulty dealing with hierarchical structures, which is a big problem when tackling linguistic processing because hierarchical structures are the foundation stones of modern linguistics (Marcus, 2018). These flaws have led many to advocate for more significant interaction between neural networks and linguistics to benefit from one another (Pater, 2019; Linzen, 2019). This thesis is an integral part of this discussion by presenting a tentative way to develop an integrated model. Linguistics is critical in this equation because linguists can contribute from all three fronts by delineating the linguistic constraints that have to be met by computational models while keeping in mind the empirical results from cognitive science (Linzen, 2019).

CHAPTER II

COGNITIVE SCIENCE AND PREDICTION

Until recently, the idea that someone would be able to predict upcoming linguistic information was viewed as improbable because it was deemed cognitively too demanding to deal with the massive space of possibilities (Kutas et al., 2011). It was only more recently that linguistic processing models shifted away from the modular view of language (Fodor, 1983) and turned to less constrained and more interactive models of language processing (Ferreira & Lowder, 2016). In turn, even though some argued that the role of prediction for actual word recognition seems to be minimal (Brouwer et al., 2012), the fact that linguistic pre-processing (anticipation or pre-activation) plays an active role in sentence comprehension is now more widely accepted than ever (Delong et al., 2014). It has become an important field of research (Jaeger & Snider, 2013; Van Petten & Luka, 2012; Levy, 2008; Nieuwland & Van Berkum, 2006).

For the most part, recent studies support the conclusion that prediction mechanisms are the best way to explain a lot of empirical results (DeLong et al., 2005; Kutas & Federmeier, 2011; Lau et al., 2013). For example, Ito et al. (2018) showed that even if anticipatory behaviors involved a higher cognitive load, it was still advantageous during linguistic processing because of the benefits of a correct prediction compared with the additional cost of a wrong prediction. Rommers &

Federmeier (2018) also argues that a verification process speeds up the processing of correctly predicted linguistic units because it allows for less thorough analysis as the brain has already partially treated that new information.

However, despite all the research on the subject, it is still difficult to grasp the exact role of prediction within language processing. One reason for that might be that a prediction can be formed from several different kinds of processes, be it linguistic pre-processing, neural pre-activation, or a simple expectancy of upcoming semantic content or syntactic structure (Delong et al., 2014; DeLong & Kutas, 2020), and it could be challenging to try and measure these processes independently.

2.1 Different kinds of predictions

Linguistic prediction mechanisms are generally separated into Type I, which is about automatic activation of associated information, and Type II, which is related to a more deliberate and reflexive mechanism (Huettig, 2015). Similarly, Brothers et al. (2017) differentiate between these two types by naming them associative pre-activation (Type I) or specific lexical predictions (Type II).

According to the first view, a given concept is associated with other concepts that are automatically passively co-activated when the first concept is processed because they share some features (Lau et al., 2013). Nieuwland (2019) gives the example of the word *chocolate* that facilitates the understanding of the word *candy* because the two are related, and hearing the first word will automatically pre-activate the second one.

In the second option, the prediction is derived from the interpretation of the meaning expressed by the words in the context (Kutas et al., 2011). The second family of views about predictions is called *active prediction* because it involves using an inferential process to derive the prediction. As opposed to passive co-activation, the active prediction is a prediction about the upcoming meaning and not only about a specific lexical item. This active prediction is a decision-like cognitive process that allows the hearer to weigh the possible options for a prediction and choose the one that best maximizes its utility (Kuperberg & Jaeger, 2016).

The distinction between the two is sometimes described as the difference between intra-lexical association (or spreading activations within the lexicon) and meaning associations (or constructed relationship between different words) (Ferreira & Lowder, 2016) or even 'dumb' prediction and 'smart' prediction Karimi et al. (2019). In this thesis, following Nieuwland (2019), I differentiate between the two kinds of predictions by comparing the two processes: the first one is a passive prediction while the second one is an active inferential process.

2.1.1 Passive pre-activation: Similarity

According to the passive pre-activation view, anticipation is not about forming a precise prediction of the next upcoming information but rather about activating information related to the processed information. For example, if we consider the lexical level, then a concept associated with other concepts would be automatically passively co-activated when the first concept is processed (Lau et al., 2013). For example, the words *windy*, *kite* and *fly* are semantically associated which means that when one hears *windy* and *fly*, the word *kite* then also becomes automatically

activated, as in Figure 2.1.

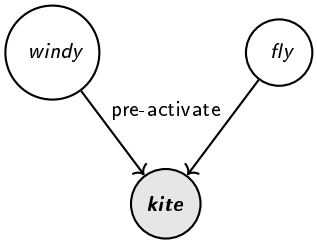


Figure 2.1 Pre-activation of the word *kite* when *windy* and *fly* are heard

This kind of anticipation is qualified as passive because the co-activation is done without any inferential input from the hearer, i.e., the activation from one word or one concept is passively spreading towards associated or related concepts. The degree to which a related word is activated depends on its relatedness with the primarily uttered word. Passive co-activation can be translated in terms of similarity measures between words. I will come back to this notion of semantic similarity in Chapter 4, but the vital thing to note here is that the more similar two words are, the higher the co-activation between those two words.

When a word is uttered, only the words closely related to this word will be activated. If many words are all co-activating the same word, then this word will become the most activated, and the hearer will be able to pick this word as the most probable lexical prediction. This is precisely what was described in the case of a truncated sentence containing *park*, *windy*, *fly* from which one can predict that the following word will most likely be *kite*. One advantage of this view is that this form of spreading activation is relatively resource-free because it does not involve any attentional or dedicated cognitive processes (Brothers et al., 2017).

2.1.2 Active Prediction: Predictability

Generally, active predictions are divided into two sub-phenomena: predictability and plausibility. The notion of predictability is contingent on the concept called cloze probability (Kuperberg & Jaeger, 2016) which we discussed in Chapter 1. Predictability is directly linked with the likelihood that a participant completes a given cloze task using a particular word (Quante et al., 2018). For example, we can recall the example (18), where the continuations were *stay* (26%), *help* (21%), *leave* (10%), *dinner* (5%) (Bloom & Fischler, 1980).

(18) The kind old man asked us to ...

For most empirical studies, the linguistic context or the sentences are so particular that more general predictability measures do not apply, so they have to conduct a small cloze task beforehand if they want to control predictability. Once they have the results for predictability, it is then possible to generate examples with high and low predictability and measure how the level of surprise influences the processing of these sentences. As discussed in Chapter 1, higher predictability is linked with faster and less effortful processing (Van Petten & Luka, 2012).

According to this view, the upcoming word predicted by a hearer would generally be the one with the highest predictability according to the situation. Even if their by-products could be the same, an active prediction based on predictability and a passive co-activation should not be confused, as they ask for different cognitive processes. They also show distinct patterns of brain activity during language comprehension (Frank & Willems, 2017).

2.1.3 Active Prediction: Plausibility

Plausibility is another kind of active prediction, but instead of being related to a cloze score, it is related to a plausibility score. In a plausibility test, the participant has to rate the plausibility of events described in a sentence (Nieuwland et al., 2019; DeLong et al., 2014; DeLong & Kutas, 2020; Quante et al., 2018). For example, in Quante et al. (2018), they asked participants to rate the plausibility of the sentence pair's meaning from 1 to 5, 5 being very plausible.

(19) a. Alice brach sich ihr Bein im Wanderurlaub.
'Alice broke her leg while hiking.'

b. Der Arzt röntge ihr Bein und legte es in einen [Gips, Rollstuhl, Vogel] für zehn Wochen.
'The doctor x-rayed her leg and put it in a [cast, wheelchair, bird] for 10 weeks.'

In (19-b), the word *cast* was deemed to be the most expected word for this utterance, while *wheelchair* and *bird* were respectively judged plausible and implausible continuations. For these tasks, rather than asking to complete the truncated sentence, the participants are asked to evaluate the sentence's plausibility as a whole. Thus, these plausibility values are not related to a particular word in the sentence but are rather associated with the plausibility of the whole composed sentence. To illustrate this distinction, we can compare the plausibility values of the two sentences in (20).

(20) a. John went to the park to fly a kite.

b. John went to the park to fly a plane

In this case, we would not be comparing the plausibility of *kite* and *plane*, but the plausibility of those two words within their respective sentence. For example, in (21), it is not that *piano* itself is plausible or not, but that "John plays the piano" in its entirety is plausible. Plausibility is thus a global measure, whereas predictability is more of a local measure.

(21) John plays the piano.

Using plausibility as an active prediction process implies that the upcoming word is predicted by maximizing the whole sentence's plausibility. This process is very different from the one involving predictability because, in the latter case, the word was chosen to best fit with the other words of the sentence, whereas in this case, the word is chosen to best fit within a sentence. For example, if one wants to predict the next word of the sentence (22) using a plausibility-driven process, one has to find the word that would make the sentence (22) as plausible as possible. So the predictive process here demands that we represent the composed sentence while the predictability process is only required to represent the meanings of the words contained in the sentence.

(22) Lucie went to the library to borrow a ...

Plausibility is linked with the 'semantic integration view,' where the predicted lexical item is interpreted and then integrated within a partial semantic representation of the incomplete sentence (Nicenboim et al., 2020). According to this view, more plausible words are easier to integrate compared with implausible words.

Some authors have considered a third case of active prediction based on what they called possibility (Quante et al., 2018; Warren & McConnell, 2007), and they illustrated the difference between a merely implausible event and an impossible event with sentences like (23).

(23) She inflated the carrots.

This sentence is not only implausible, but it is semantically non-well-formed.

> Lexical knowledge indicates that the verb *inflated* requires an inflatable object because the context restricts *carrots* to refer to non-inflatable vegetables and because carrot-inflating events are never encountered" (Warren & McConnell, 2007, p.1)

In the literature, these infelicitous cases are referred to as *impossible*. In this thesis, I only consider plausibility, and I do not further discuss possibility because I treat the latter as an extreme case of the former, i.e., if it is impossible, then it is not plausible, and, if it is plausible, then it is possible. The gradability of the possibility scale lies on the same scale as plausibility itself, as it represents the lowest spectrum of cases on the scale of plausibility. In other words, it is when the plausibility becomes very low that we might transition from possibility to impossibility, and I thus conceptualize the possibility scale as being a sub-scale of the plausibility scale.

2.2 Measuring a Prediction

In Chapter 1, I briefly discussed some empirical results supporting an incremental view of linguistic processing, in this chapter I turn more specifically to empirical measures involving similarity, predictability and plausibility. There are several empirical methods to measure the effect of predictive behaviors in language processing. According to Ferreira & Lowder (2016), we have two empirical disciplines that are interested in measuring markers of linguistic prediction: psycholinguistics studies and neurolinguistics studies. The latter being interested in brain signals while the former is interested in online measures of a linguistic process like reaction times and eye-movements.[13] Both fields of research measure what Armeni et al. (2017) calls covert linguistic markers. These are not direct measures of prediction, as they rather are indirect measures of wrongful predictions. Instead of determining the lexical prediction of a reader/listener in a given linguistic context, covert markers measure the effect of a prediction, mainly the consequence an unexpected word has on linguistic processing. In this section, I start by discussing studies about eye-tracking experiments and results that link reading time to the role of prediction within language processing and then discuss in detail ERP studies about the N400 component and other indirect measures of the effect of prediction on language processing.

---

[13]In Armeni et al. (2017), they used the broader term of Cognitive neuroscience, but throughout this thesis, I use the term neurolinguistics to refer to cognitive neuroscience studies specifically interested in linguistic prediction and language processing.

2.2.1 Psycholinguistics

Eye-tracking studies measure prediction through reading times and direction of gaze, and they have shown that the predictability of a word strongly influences the time it takes to read that word, and higher predictability has been linked with more efficient language processing (Staub, 2015). Language processing is thus facilitated when upcoming inputs are correctly anticipated.

A measure that has been widely used in psycholinguistics is the reading time (RT). Studies about prediction and reading time have supported the idea that prediction plays a role in language processing (Levy & Gayler, 2008; Linzen & Jaeger, 2016), and the amount of time it takes to read a word has been linked directly with processing difficulties. In turn, processing difficulties have also been correlated with the ability one has to predict the upcoming word (Smith & Levy, 2011, 2013; Staub et al., 2015). The correlation between predictability and cognitive effort makes sense if we assume that the cognitive representations for an expected word are more activated than those for a less expected word. This higher activation implies it is easier to retrieve the word from memory and process it (Roland et al., 2012).

On the other hand, whenever a new word is processed, it shifts the expected structure of the sentence. Furthermore, when this shift in expectation is significant, it takes longer to process the sentence, and this results in a longer reading time (Kutas et al., 2011). A garden-path sentence like (24) is a good example of this kind of shift in the expected representation of a sentence.

(24) The horse raced past the barn fell.

Levy (2008) was one of the first to use the correlation between reading time and processing difficulty to show that unpredictable sentence structure took longer to read. Levy then went further by describing this unexpected (or unpredicted) shift in sentence structure as a surprisal effect which is a measure of the transition between different informational states when new words are processed. The surprisal is the unexpectedness of a word given the context, and it is mathematically represented by taking the logarithm of the probability of the predicted difficulty over processing the ith word (w_i) given the context and the current input ($w_{1...i-1}$) (Levy, 2008), as in (25).

(25) $\text{difficulty} \propto -\log P(w_i|w_{1...i-1}, \text{CONTEXT})$

This probability is the probability that a given word is expressed given a particular linguistic context. In other words, when a word is least expected, it takes more cognitive energy to process it. This correlation between processing difficulty and conditional probability has been empirically supported by results from reading times studies (Smith & Levy, 2013), from eye-tracking data (Demberg & Keller, 2008) and from ERP studies (Frank et al., 2015). Even if some recent results question the logarithmic nature of this relationship and instead argue in favor of a linear pre-activation account (Brothers & Kuperberg, 2021), surprisal is still used widely when modeling sentence processing (Futrell et al., 2020; Venhuizen et al., 2019).[14]

It is possible to have different interpretations of the effect of surprisal (Roland

[14]The notion of linear relationship might be a bit confusing in the literature about surprisal because some are referring to the relationship between surprisal itself and processing difficulty (Ryskin et al., 2020; Goodkind & Bicknell, 2018) which is linear, while others refer to the relationship between predictability and processing difficulty (Brothers & Kuperberg, 2021) which is logarithmic according to the surprisal account.

et al., 2012). For example, Jurafsky (2002) considers surprisal to be related to the amount of information conveyed by a word. In contrast, Hale (2001) characterizes it as the probability strength of the interpretations rejected once a word has been processed. The interpretation that was first proposed by Levy (2008) is most commonly viewed as the relative difference between the probability that the world is in a given state just before reading this new word and just after. In other words, the more surprising is the word, the more the posterior probability will be different from the prior one.

This strict correlation between predictability and RT has lately been challenged because recent empirical studies showed that in cases where the predictability is low, reading times are only faster when the previous linguistic context is highly constraining (Staub et al., 2015). For example, (26-a) is said to have low predictability with a high level of constraint, and *election* has a high predictability score, 61%. Although they have lower cloze scores, all the other responses were nevertheless related in meanings with *election*: contest, battle, award, and prize (Staub et al., 2015). On the other hand, (26-b) has low predictability and low constraining as it elicited very diverse response such as *bird*, *girl*, *cat*, and *crying*.

(26) a. [Low predictability, High constraining]: He complained to win the ...

b. [Low predictability, Low constraining]: He heard a faint sound of one ...

In their experiments, Staub et al. (2015) asked participants to verbally complete sentences while recording their response time, and they measured shorter times when predictability was high. However, when predictability was low, responses were only faster when the linguistic context was highly constraining, i.e., only for cases like (26-a). These results point toward a redefinition of the strict relationship

between cloze score and predictability because it seems other factors might also intervene in the process. These results imply that predictability cannot be the only factor influencing RT measures (Yun et al., 2012).

Roland et al. (2012) has also observed an effect of similarity on reading times. They showed that words similar to the previous linguistic context were processed faster, and they found this effect to be independent of the influence of the predictability score. The similarity in this scenario is defined as the semantic similarity measure described in the passive pre-activation section. To illustrate their results, let's consider (27-a) and (27-b), a modified version of their examples.

(27) a. The soldier jabbed the angry lion with a/an ...
b. The soldier attacked the angry lion with a/an ...

In the context of (27-a), the next word could be *spear*, *sword* or *machete*, whereas in the context of (27-b) it could be *sword*, *stick*, *knife* or *rock* or *gun*. According to Roland et al. (2012), in order to find the word that is the best candidate in its respective context, the interpreter might use a combination of both the notion of predictability and the measure of similarity, and, most importantly, they argued that the similarity effect between these words and the preceding context would supersede the predictability effect for the same words. In Figure 2.2 (Roland et al., 2012, adapted from their Figure 2), we can see that the similarity between the candidates is better in the *jab* context than in the *attack* context since both *rock* and *gun* are less related with the other possible continuations. What Roland et al. (2012) showed is that the reading time for cases where the similarity was bigger between potential candidates was shorter than for cases where the similarity was less significant, independently of the predictability values of the candidates.

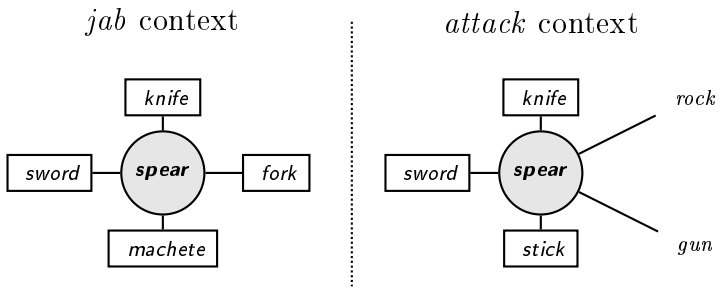


Figure 2.2 Representations of the semantic similarity between *spear* and other options for both the 'jab' and the 'attack' context

Their results thus support the idea of this precedence of the similarity over the predictability. It is interesting to note that despite their results, Roland et al. (2012) did not dismiss predictability altogether, and they opened the door to the integration of the two effects into a hybrid model of RT. This conclusion is important because it contradicts the model of Levy (2008) which is based strictly on predictability. An influence of similarity on reading time has also been reported by Yun et al. (2012), which supports the idea that RT is not only influenced by predictability but that it might also be dependent on the similarity between words.

In another experiment, Roland et al. (2012) measured another influence on the RT, but this time it was an interaction between frequency and predictability. In psycholinguistics, the notion of frequency is tied with the unconditional probability that a word is occurring, regardless of the specific linguistic context (Smith & Levy, 2013). This result, combined with the previous one about similarity, only strengthens the assumption that predictability does not act alone.

Finally, the last point I want to mention regarding predictability is that even though the correlation between predictability and reading time has been established and accepted (Staub et al., 2015; Smith & Levy, 2013) and even tough we are able to calculate predictability and cloze probability, there is a an open ques-

tion about what these measures represent in terms of cognition (Smith & Levy, 2011).

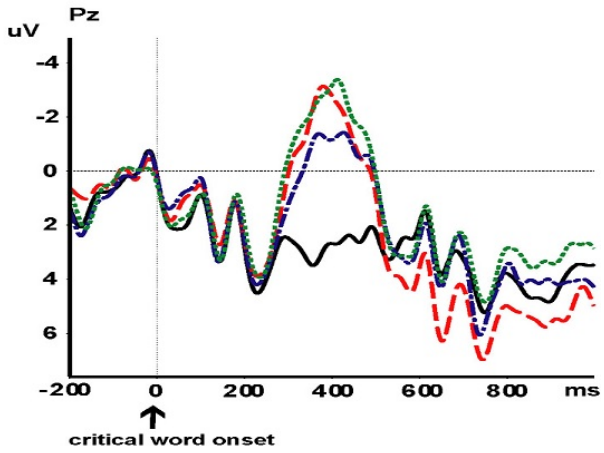
The difficulty here lies in the many possible ways to interpret RT results because of the different ways to theorize the underlying process. One perspective state that the predictability effect comes from the anticipatory pre-activation of words, i.e., similarity, and another stance supports the idea that the effect is happening only post hoc, i.e., that it is happening when the word meaning is integrated (Smith & Levy, 2013). A third option would be for someone to predict the next upcoming word using a bit of both process (Staub et al., 2015). In Chapter 4, we develop a hybrid view where similarity and plausibility each have their role to play during the derivation of a linguistic prediction.

2.2.2 Neurolinguistics

In recent years, many event-related potential (ERP) studies focussed on prediction, and it is fair to say that the N400 component has become central in the neurolinguistic literature about prediction. The N400 component is a negative deflection in an ERP waveform peaking around 400 ms after stimulus onset and is more significant over centro-parietal electrodes (Cosentino et al., 2017). If in reading time studies, unpredicted information was correlated with surprisal and longer reading time, in ERP experiments, unexpected information is correlated with a larger N400 effect. The N400 effect is used to show that lower surprisal is linked with predictable words, and faster and effortless processing (Kutas et al., 2011; Van Petten & Luka, 2012).[15] To illustrate this N400 effect, we can look at

[15]See Nieuwland et al. (2018); Nieuwland (2019) for an exhaustive review of N400 results.

Figure 2.3 (Ito et al., 2016, adapted from their Figure 1) where the lowest value is for the most expected word, namely *book*. According to this graph, the second most expected word is *page*. The words *hook* and *sofa* elicited the higher N400 effect, and they are thus deemed to be the most surprising of the four.



The student is going to the library to borrow a hook/page/sofa/book tomorrow.

Figure 2.3 Illustration of N400 effect

The N400 effect is the difference in N400 components measured under two different stimuli. This effect was described for the first time by Kutas & Hillyard (1980), and it has been extensively investigated ever since (Nieuwland et al., 2018; Nieuwland, 2019). In their original experiment, Kutas & Hillyard (1980) compared N400 components for words that had mismatches for meanings compared with the first part of a sentence. For example, in (28), the largest N400 component should be the one for the word *socks* because it is unexpected, and it thus requires a sizeable cognitive effort to process (Kutas et al., 2011).

(28) He spread his warm bread with butter/socks.

N400 effects are also observed when a visual stimulus or an image is unexpected in a given situation (Lau et al., 2013), and, in this sense, it transcends linguistic processing as a measure of the effect of unexpectedness, i.e., the effect of a wrong prediction.[16]

In Chapter 1, I mentioned that I would not discuss in length linguistic prediction below the word level, but it is worth mentioning that ERP studies about gender prediction have tested the effect of phonological information on the N400 component. In Ito et al. (2020), they compared expected nouns that had different phonological or gender classes. The truncated sentences like (29) were continued either by an expected article and expected noun, a phonological mismatch article and a noun, or a gender mismatch article and a noun.

(29) Il traffico in autostrada e rimasto bloccato a
The traffic on the motorway came to a standstill
causa di...
because of...

a. $un_{masc.}$ $incidente_{masc.}$ *(accident)*

b. $un_{masc.}$ $scontro_{masc.}$ *(collision)*

c. $un_{fem.}$ $inodazione_{fem.}$ *(flooding)*

Ito et al. (2020) showed that gender information was processed quicker than phonological information, although both kinds of mismatches elicited a larger N400 effect. Similar studies were also performed in other language and they gave rise to similar results, e.g. Szewczyk & Schriefers (2013) (Polish), Nicenboim et al. (2020) (German), Van Berkum et al. (2005) (Dutch), and DeLong et al. (2005)

[16] I do not discuss these extra-linguistic scenarios in this thesis, but the fact that N400 effect also applies to visual expectations should not be a surprise for those who conceptualize language interpretation as being linked with perceptual information.

(English). The exact nature of this N400 effect is still debated, but most agree it is certainly related to the degree by which input matches its predicted value (Ito et al., 2016) or related to the "ease of accessing semantic features and properties associated with incoming words" (Kuperberg, 2016, p.603).

The last thing to note about the N400 results is their empirical role in helping to differentiate different processing streams of incoming information that are taken into account when deriving the prediction of the upcoming word (Kuperberg et al., 2020).

Apart from the N400 components, there is also a late positive response named the P600 (Leckey & Federmeier, 2020). Contrary to the N400, which is measured between 300-500ms after the onset of the target word, the P600 is measured between 600-1000ms after this onset (Brouwer et al., 2017; Sassenhagen & Fiebach, 2019; Kim et al., 2018; Bornkessel-Schlesewsky & Schlesewsky, 2008). In a series of experiments where they manipulated the length of the prior context, Brothers et al. (2020) compared the N400 and P600 components' responses. They observed that in a setting where the prior context was absent, it did not elicit a P600 response. Whereas they observed a P600 in cases the prior contexts were constraining. More specifically, they observed a *late posterior positivity* for locally constraining contexts, and both a *late posterior positivity* and a *late frontal frontal positivity* for globally constraining contexts as in (30).

(30) a. No context:
James unlocked the (door/laptop/gardener).

b. Locally constraining context:
He was thinking about what needed to be done on his way home. He finally arrived. James unlocked the (door/laptop/gardener).

c. Globally constraining context:

> Tim enjoyed baking apple pie for his family. He had just finished mixing the ingredients for the crust. To proceed, he flattened the (dough/foil/onlookers)...

In the first experiment, Brothers et al. (2020) did not use any context as in a conventional cloze task. In the second and third experiments, they added some context in front of the truncated sentence. In (30-b), the events described by the first two sentences are not strongly associated with the event described in the third sentence, and it did not help select the best candidate since *door* and *laptop* are both possible. Finally, for cases like (30-c), the event described in the third sentence followed naturally from the interconnected events described in the first two sentences. Their findings point toward the importance of contextual constraints in predicting the upcoming word. However, most importantly, they support the idea that we could differentiate ERP components in terms of their respective contribution during linguistic processing.[17]

2.3 Cognitive Constraints for a Model of Linguistic Prediction

In this section, I look more closely at some key results about the way linguistic prediction is realized cognitively. I am interested in underlining empirical results that help us retrieve the structural and cognitive constraints we must consider when developing a model of linguistic prediction.

---

[17] In a different study, Brothers et al. (2017) showed that the broader linguistic context and the specific goal of the listener could influence the anticipatory process during language processing. This other result adds weight to the hypothesis that the context of the enunciation has to be taken into account when trying to understand how lexical prediction operates.

The most important thing to note from the many empirical results on linguistic prediction is that there are no hard architectural constraints on the flow of activity within the interpreter's mind and both influences coming directly from the input, and predictions derived from information that is already known, are possible concurrently (Kuperberg, 2016). As put forward by Khachatryan et al. (2018), the N400 component reflects the interaction between these two processes. When the discrepancy between the two is more significant, i.e., when the input does not correspond to the prediction, the ERP responses become larger, and this signals a wrong prediction or an unexpected word.

2.3.1 Distinguishing the kind of prediction

The correlation between predictability, plausibility, and similarity makes it hard to clearly understand their respective effects on linguistic processing because it is difficult to differentiate between their contribution (see section 2.1). This correlation is easily understandable because when an N400 effect is encountered in a high-cloze situation, i.e., a high-predictability context, it might also be the case that we also are in a high-similarity or a high-plausibility scenario. For example, DeLong et al. (2019) and Ito et al. (2016) noticed that participants' plausibility values were higher when semantically related words were involved, even for cases when these words did not correspond to the given context of the sentence.

Additionally, following Nieuwland et al. (2019), if the truncated sentence is (31), then the word *bicycle* is more predictable than the word *elephant*, but it is also the case that the word *bicycle* makes the whole sentence more plausible compared with a sentence containing the word *elephant*. In this particular case, predictability is

related to higher plausibility, making it difficult to single out any of those two contributors to the N400 effect.

(31) You never forget how to ride a...

Simultaneously, the word *bicycle* might be predicted from an active process using a cloze score, i.e., predictability, or it might also come from a passive co-activation process because it co-occurred with the word *ride*. The same N400 results could thus be caused by a mixture of the three kinds of predictions discussed previously, and it is of utmost importance to use experimental conditions specifically designed to help us differentiate between these potential contributors. Nieuwland (2019) explored these potential cross-contributions from existing studies and performed their own set of experiments by varying the nature of the three variables (Nieuwland et al., 2019). Their experiment used 80 different sentences, each having two conditions: expected and unexpected article noun combinations. They also measure predictability, plausibility, and similarity distinctively for every article noun combination. Predictability was given as a cloze score going from 0 to 100%, plausibility was measured on a scale from 1 to 7, and similarity was measured using Latent Semantic Analysis (LSA). I will describe in length how it is possible to measure similarity in Chapter 4, but, for now, we can still illustrate the possible variations between these three conditions using these examples taken from Nieuwland et al. (2019).

(32) It was difficult to understand the foreign professor because he had an/a ...

a. *accent*: Predictability of 100%, Plausibility of 6.13, Similarity of 0.29
b. *lisp*: Predictability of 7%, Plausibility of 6, Similarity of 0.12

(33) Although the basketball team's defense was very strong they did not have so much of an/a ...

a. *offense*: Predictability of 34%, Plausibility of 5.32, Similarity of 0.26

b. *coach*: Predictability of 0%, Plausibility of 3.33, Similarity of 0.53

(34) Every time they went for walks Sylvia's dog Rex would break into a run as soon as he spotted a/an ...

a. *cat*: Predictability of 27%, Plausibility of 6.1, Similarity of 0.34

b. *owl*: Predictability of 10%, Plausibility of 4.3, Similarity of 0.29

In (32), it is the predictability that is the defining contribution since the plausibility and the similarity are not varying significantly between the two continuations. In (33), is it the plausibility and the similarity that contribute more (taking into account that both *offense* and *coach* have pretty low predictability scores), and in (34), it is the plausibility that has the most considerable variability. They measured the N400 components for all these cases, and their results were consistent with a hybrid, multi-process view of prediction where predictability, plausibility, and similarity each contribute to the lexical prediction that was formed by the hearer (Nieuwland et al., 2019).

Veldre & Andrews (2018) arrived at a similar conclusion while using eye-gaze recordings and target words that were either plausible or predictable. They showed that plausibility affected target fixation duration measures that could not be explained by relying solely on predictability scores.

(35) She put firewood into the... (stove)

In another study Khachatryan et al. (2018) used intracranial EEG measures to

argue that the complex interaction between the three kinds of prediction depends on the specific task and local context of the target sentence. For example, in highly constraining cases like (35), the linguistic prediction seems to be mainly derived from predictability, and conversely, when the sentence is not constraining enough, the contribution from the lexical associations is required.

2.3.2 Multiple Processing Streams

In order to model word-level linguistic prediction, not only do we have to be careful about these three contributors, but we also have to take into account the different potential sources of linguistic information that could simultaneously influence the derivation of this prediction. When tuned correctly, ERP results can help understand these different mechanisms contributing to linguistic prediction. Until a couple of years ago, the standard view was that the N400 component was related to semantic surprisal, while the P600 component measured the syntactic surprisal Osterhout & Holcomb (1992); Hagoort et al. (1993). The two were also usually thought of as being co-dependant effects, but when a more thorough examination was done using more specific stimuli, it was shown that the two components could be independent of one another (Kuperberg, 2007; Brouwer et al., 2017).

Payne et al. (2015) found that the N400 component was influenced by the degree of syntactic constraint coming from the context, and Brothers et al. (2020) provided evidence that the P600 component, in some contexts, was not solely a measure of syntactic violations. To explain these results, Kuperberg (2007) posited two processing streams that could both influence ERP measures. In (36-a), the object noun *tourist* is both related with the overall topic, and respects the animacy

constraints imposed by the verb *told*, whereas in (36-b) the animacy constraints are violated by the object *suitcase*.

(36) a. The woman told the tourist ...

b. The woman told the suitcase ...

By comparing ERP measures for when a preceding discourse was added or not, Nieuwland & Van Berkum (2005) got a large P600 effect but no N400 effect in the case of (36-b) without any preceding discourse, but they observed an N400 and no P600 effects when a preceding discourse was taken into account. Kuperberg (2007) explained these results by arguing for dissociating syntactic and semantic processing into two separate streams. These results suggested that these two streams are highly interactive, but also that they can act independently (Kuperberg, 2007, p.41).

According to Kuperberg (2007), semantic-thematic representation is a hybrid process that is dependent on the syntactic stream for the combinatorial of lexical items together, but it also needs to be interpreted by a semantic stream. In other words, the semantic-thematic relations are derived by the syntactic streams, but their plausibility is assessed by their relationship with the real world (Kuperberg, 2007). It is from the conflict between these two processing streams that the P600 would arise (Kuperberg, 2007).[18]

From there, it seems clear that when discussing cloze task or linguistic prediction, we must consider two independent processing streams: the syntactic and the se-

[18]In a recent paper Fedorenko et al. (2020) showed fMRI results that do not support this independence between the syntactic and lexico-semantic processing, but their results still support the prevalence of the latter over the former. These results are vital if we want to build a realistic mechanistic model of linguistic processing, and they are still compatible with the model I present in Chapter 4 and Chapter 5.

mantic one (Kuperberg et al., 2020). Moreover, the P600 component has then been further divided into two other sub-effects: a semantic verb-argument violation and a syntactic violation (Kuperberg, 2007; Shetreet et al., 2019). These two P600 components, i.e., the syntactic P600 and the semantic P600, are considered to be strong evidence that syntactic and semantic processing can be independent. There are ways of explaining these with single streams models (Brouwer et al., 2017), but for the purpose of this thesis, we follow the dominant view that these streams are processed independently.

The first stream is responsible for processing the structure of linguistic input via combinatorial mechanisms based on morpho-syntactic rules, and the second stream is about semantic-memory retrieval of lexical items.

The semantic and syntactic streams appear interdependently modulated even though they could still be dissociable (Kuperberg, 2007). To illustrate this, we can use these examples from Chow et al. (2016b):

(37) The restaurant owner forgot which customer/waitress the waitress had served during dinner yesterday.

(38) The superintendent overheard which tenant/realtor the landlord had evicted at the end of May.

In their results, Chow et al. (2016b) showed the N400 component measured in (38) on the verb *evicted* was smaller for sentences containing *tenant* relative to the one containing *realtor*, but they showed no difference between the two possible sentences in (37). This result tells us that the interpreter did not use the syntactic information from the sentence to predict the upcoming verb *served*. To explain this result, Chow et al. (2016a) proposed separate processing streams:

one mechanism responsible for pre-activating verbs that are thematically related to the arguments without considering the structural roles of these arguments, and one slower syntactic mechanism that assigns lexical items to structural, thematic roles. This interaction between the two streams depends on the evidence available at the time of the input integration (Kuperberg, 2016).

Having two different processing streams is also compatible with evidence about parafoveal processing (Veldre & Andrews, 2018). Parafoveal processing is related to how a reader previews an upcoming word using his gaze's parafoveal area. In doing so, it should facilitate the subsequent identification of this word. In their experiment, Veldre & Andrews (2018) replaced a target word within a sentence with a different word as soon as the reader made a saccade over the word that preceded the target word, i.e., as soon as the gaze crossed the word 'spare' in (39). To understand the influence of syntactic and semantic information in reading, they compared the fixation duration between four kinds of continuations: a continuation identical to the target words (39-a), a plausible continuation that shares the same grammatical class as the target word (39-b), an implausible continuation that shared the same grammatical class as the target word (39-c), and an implausible continuation that was from a different word class to the target word.

(39) She eventually found a spare [...] behind the crowded bar.

a. stool
b. glass
c. uncle
d. begin

From their results, Veldre & Andrews (2018) concluded that parafoveal processing

was most beneficial when both syntactic and semantic information was plausible, which suggests that readers can treat both kinds of information in parallel.

One last thing we should note about these two processing streams is the precedence of the semantic stream over the syntactic stream. Precedence not because the syntactic stream is contributing less, but because in the absence of any syntactic indication, it is still possible to predict an upcoming word from semantic information only, while the contrary is not possible (Baggio, 2018). This apparent limitation of the syntactic processing stream is in line with the idea that semantics is autonomous from syntax and can supersede it if necessary. In other words, semantics generally has a more significant effect than syntax in prediction tasks (Kuperberg, 2007). According to this view, semantics is not only autonomous from syntax, but it is also prevalent to syntax (Michalon & Baggio, 2019; Baggio, 2018; Gärdenfors, 2014; Morgan et al., 2020). The only time syntax has a more significant influence on the prediction in cases where the semantic evidence is weak (Kuperberg, 2016), as in the case of 'nonce words' as in (40).

(40) The griop surked the ...

This precedence of semantic over syntactic processing does not contradict ERP results where the P600 components had been triggered by a syntactic anomaly. However, it does imply that the syntactic stream is not strong enough by itself to form a prediction for the next word.[19] The fact that both semantic memory-based associations and semantic-theme relationships could in practice overcome the syntactic stream makes for a less syntactico-centric processing perspective

[19]See Kim et al. (2015) and Kuperberg (2016) for different examples where semantically incongruous sentences were repaired by ignoring the syntax so that the correct thematic roles could be re-attributed accordingly and Kuperberg (2007) for a list of conditions where such semantic precedence could happen.

than what is usually considered (Kuperberg, 2007).[20] This view is also in line with the parallel processing approach of Jackendoff (2007); Culicover & Jackendoff (2006) which states that language components are processed distinctively and are put together using different interface rules that take place between the different streams.[21]

2.3.3 Representational Level and Linguistic Prediction

The idea of the lexical level being the level that is predicted is in line with the word recognition hypothesis (Nieuwland, 2019). According to this hypothesis, the prediction is measured against the real input only after the specific word form has been recognized, i.e., after the lexical item has been processed. This idea is in opposition to the sensory hypothesis for which predictions are implemented as perceptual templates representing both the phonology and the visual appearance of a word (Nieuwland, 2019).

The question of the time measurements of the prediction, i.e., the temporality of the N400 surprisal effect, is essential to differentiate between these two views. This distinction originates from the difference in the trigger of surprisal: pre-lexical, lexical, or post-lexical. The pre-lexical phase usually refers to phonological predictions, the lexical phase to predictions about the word form, and the post-

---

[20] This idea has also been discussed and argued for in the 'good-enough' approach of linguistic processing (Ferreira & Swets, 2002; Ferreira & Lowder, 2016; Ferreira & Chantavarin, 2018) which is presented in Chapter 5.

[21] Culicover & Jackendoff (2006); Jackendoff (2007) proposed to treat semantics as autonomous to syntax. Under this view, from passive elements of syntactic trees, words become active in determining the structure of an interpretation at the phonological, syntactical, and semantic levels.

lexical phase refers to the moment when the meaning expressed by a given word is integrated within a sentence-level representation (Nieuwland, 2019; Kutas et al., 2011).

At the pre-lexical stage, non-semantic lexical features like grammatical and phonological features might also be pre-activated, but it does not imply that the word form is necessarily also pre-activated (Baggio, 2018). To illustrate this difference, we can look at Figure 2.4, where the surprisal effect for an incorrect prediction is measured from the onset of the word in the case of pre-lexical anticipation, whereas it is measured after the meaning of the word had been interpreted in the case of post-lexical anticipation. Regarding the levels of representations: at the pre-lexical phase, we are dealing with representations that are below word-level, while at the post-lexical phase, we have to use the above word-level representations to compose the meaning of the words together.

Figure 2.4 Different triggers of prediction: pre-lexical, lexical, post-lexical

These three different processing stages are linked with the distinction between the 'access view' and the 'semantic integration view' of linguistic prediction (Nicenboim et al., 2020). According to the 'access view,' every time one reads some word w, it triggers memory access, and all these memory accesses are combined to pre-activate the next word. The interesting thing with this view is that this pre-activation could be happening both at the pre-lexical and the lexical level. In other words, when we read a word, not only does it pre-activate other lexical

items that are associated with it in our memory, but it also pre-activates semantic features that are themselves linked with lexical item (Kutas et al., 2011; Rabovsky & McRae, 2014; Kuperberg & Jaeger, 2016).[22]

As for the 'semantic integration view,' it states that after reading the word w, the reader integrates it within the local and global contexts expressed by the sentence (Baggio & Hagoort, 2011). When word w is encountered, the semantic representation is partial because the sentence is still incomplete at this point (Nicenboim et al., 2020). ERP measures would thus be related to the easiness of integrating a new word within this partial representation. For example, if the word is easily integrated because it is easily predictable, then the N400 component would be lower.

It remains unspecified whether the N400 results can be directly linked with a compositional process like the 'semantic integration view' or just a co-activation of meanings in the sentence like the 'access view' because we seem to have evidence supporting both sides (Nieuwland et al., 2019). For example, using a Bayesian random-effects meta-analysis on publicly available data, Nicenboim et al. (2020) were able to show clear evidence supporting the 'access view' account, whereas Fitzsimmons & Drieghe (2013) data about word skipping during reading were compatible with an approach of linguistic processing where every word that is read is integrated quite rapidly within a sentential representation. However, the critical thing to note is that these two views about lexical processing units' levels are not incompatible. In other words, if we want to build a model of linguistic prediction, we ought to take both the pre-lexical semantic features and the post-lexical semantic information into account.

---

[22]It could as well pre-activate other kinds of features like grammatical and phonological features (Nicenboim et al., 2020), but these are not discussed in this thesis.

According to Nieuwland (2019), one thing that ERP results are showing is the necessary departure from the strong prediction view about the role of anticipation in linguistic processing. According to this strong prediction view, a prediction formed by a hearer is represented at all levels at the same time, which means that the anticipation is not limited to the lexical meaning but also includes the phonological and grammatical pre-activation associated with the predicted word. Following this strong prediction view, when someone is in an unexpected situation, the surprisal effect should follow not only from the meaning of the word but also from the phonological activation coming from the input. In other words, if the first syllable of the following word does not correspond to one of the predicted words, we should then observe an effect similar to the N400 component but happening a bit earlier. The problem is that we do not have, at this point, enough evidence that the initial phoneme of the predicted noun is pre-activated at the same time as the word itself, and we have "no clear evidence to support routine probabilistic pre-activation of a noun's phonological form during sentence comprehension " (Nieuwland et al., 2018, p.14).[23]

2.3.3.1 Semantic Features

Following Kuperberg (2007), we can distinguish between two levels of representations: the lexical level, which is about the relationships between words stored in the semantic memory, and the feature level, which is about the relationships between features related to these words. These features include thematic relation-

[23]Prediction at the phonological level has been associated to an N200 effect detected 200 ms after the stimulus's onset (Boudewyn et al., 2015; Connolly & Phillips, 1994), but, as argued by Nieuwland (2019), the N200 results by themselves are not strong enough to validate the strong prediction view.

ships that can constrain the number and types of arguments assigned by a verb (Kuperberg, 2007), and they usually sit in between the syntactic and semantic streams.

(41) The soldier jabbed the angry lion with ...

To illustrate the role of such features, we can go back to the example from Roland et al. (2012) where (41) had to be completed. In this case, when considering only the lexical level, *machete* would be a potentially strong candidate. However, when we take into account the lexical features of the word *jabbed* which is associated with "sharp pointy object," it activates another subset of the lexicon, and this activated subset might compete with the lexical level so that *machete* might not have the highest cloze score after all. The complete activation pattern of the words present in the lexicon is derived from the contribution coming from both kinds of representational levels: lexical and semantic features.

There has been much ERP evidence supporting the view that semantic features facilitate linguistic prediction (Boudewyn et al., 2015; Kuperberg, 2013). For example, Federmeier & Kutas (1999) showed the N400 component was lower for words that had the same semantic category as the most expected candidate. If we take the sentence in (42) as given, (42-b) elicited a lower N400 than (42-c) because both 'pines' and 'palms' are trees, while 'tulips' is not.

(42) They wanted the hotel to look more like a tropical resort. Along the driveway, they planted ...

a. palms

b. pines

c. tulips

The use of semantic features dates back to studies on categorization and conceptual representation (McRae et al., 2005), but they are now used frequently in computational models of linguistic prediction (Nicenboim et al., 2020). For example, Rabovsky (2019) has used them to simulate N400 results for a series of different cases (Rabovsky & McRae, 2014; Rabovsky et al., 2016, 2018; Rabovsky, 2020).[24]

2.3.3.2 Semantic Information above word-level

According to the post-lexical hypothesis, the N400 effect caused by the mismatch between the prediction and the actual input is triggered only once the meaning of the new word (the input) is integrated within a sentence-level representation of the meaning expressed by the previous words. In other words, it would not directly be the word itself that is predicted but the meaning associated with it. Following this view, sentence-level representations would be at the center of predictive behavior.[25] At this post-lexical stage, once this above word-level representation is derived, then it can be used to generate a prediction about the word that would best fit the meaning expressed by the sentence-level representation, i.e., much like the plausibility type of prediction.

A significant development towards understanding these above word-level influences is the multiple-layer representational model of Kuperberg et al. (2020) in

---

[24]I discuss the semantic features and their role in linguistic prediction in more detail in Chapter 4.

[25]This is very similar to what is described by Rabovsky (2019); Rabovsky et al. (2016, 2018); Rabovsky (2020) as a Sentence-Gestalt (SG) representation. I discuss their model in Chapter 6.

which they present a theory of linguistic prediction involving three hierarchical levels of representations. At the top of the hierarchy, the situation model is a model of the more global situation surrounding the event described by the sentence. This level represents the structure of the events and comprises the complete set of events, actions, and characters that could be involved in them. In the middle level, the representation of the event itself contains the information required to tell which sets of events are compatible with the situation described in the sentence. Finally, the lowest level of representation is the semantic level, which is concerned with the semantic features of different individual words associated with these two other levels (Kuperberg et al., 2020). Each of these three representational levels plays a role when trying to complete an utterance. As an illustration, consider (43).

(43) a. [Preceding context]: The lifeguards received a report of sharks right near the beach. Their immediate concern was to prevent any incidents in the sea.

b. Hence, they cautioned the ... (swimmers/trainees/drawers).

Figure 2.5 (Kuperberg et al., 2020, Figure 1) depicts this hierarchical model, where, at the situation level, we have a beach scene in which we have two lifeguards, a person, and where someone is being cautioned. The event level is a representation of the event structure (e.g., the lifeguards cautioned someone), and then the semantic feature level is at the level of individual words and their features (e.g., the cautionee, or the patient, which, in this case, must be sentient and be able to move).[26]

[26]See Kuperberg et al. (2020) for a more thorough presentation of the three levels.

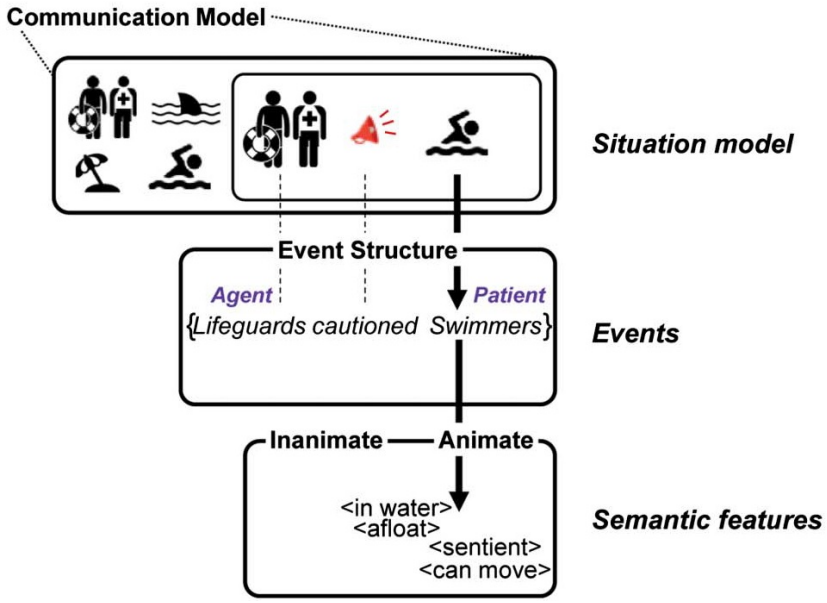


Figure 2.5 Kuperberg et al.'s hierarchical model

In their experiments, Kuperberg et al. (2020) showed that the late frontal positivity was related to updates located at the situation-level of the model. In contrast, the late posterior positivity resulted from an impossibility to update the situation level due to a conflicting constraint.[27]

The main improvement coming from this multiple-layer model over the single-layer one is that in addition to the input from the sentence itself, we now need to consider the potential interactions between the different hierarchical levels of representations because they play a constraining role in predicting an upcoming word.[28] The interactions between these representational levels are discussed in

[27]Brothers et al. (2020) findings in their first experiments seem to contradict this idea of having a higher comprehension of the context of a sentence, but it was argued that maybe the participants were not able to build such situation models because the material was not correctly contextualized.

[28]This multi-layered model is compatible with the view that hierarchical syntactic structure also conditions word-by-word expectations (Brennan & Hale, 2019).

detail in Chapter 4 and Chapter 5.

2.4 Architecture of Linguistic Prediction

So far in this chapter, I have presented three main conclusions that can be drawn from empirical results regarding linguistic prediction: the semantic stream has precedence over the syntactic stream, the predictability, the similarity, and the plausibility are all inter-related, and, finally, when deriving a prediction we have to consider representations above and beyond the level of the word. When developing a theoretically-driven model of linguistic prediction, we thus have to consider these conclusions and use them as primary cognitive constraints.

Baggio (2018) describes empirical results and introduces a processing architecture that corresponds to recent empirical evidence. The model of linguistic prediction I present in this thesis was developed with the general architecture of Baggio's model in mind, and it is thus relevant to introduce it briefly.[29]

According to Baggio (2018), linguistic processing can be divided into three systems or neuro-cognitive representations: the R-system, the I-system, and the E-system. These three systems are motivated by empirical results in neurolinguistics and psycholinguistics, and the differences between these three are well motivated in his book.

The R-system maps the meaning of the lexical items stored in the memory into relational structures constituted from these lexical units. The I-system is the interpretation system that takes care of the referential, elaborative, and inferential

[29]Please see Baggio (2018) for more detail about his approach and his conclusions.

processes at each interpretation point. It is responsible for computing a minimal discourse model, i.e., interpreting tokens for the relational structure that the R-system has generated. Finally, the E-system is the internalization of pragmatic representation, "it is a dynamic record of the status of coordination and communication with other agents" (Baggio, 2018, p.203). Its main goal is to manage the coordination between uses of a linguistic expression and its meaning with respect to the state of the world.

As far as the processing streams we just described, both the R-system and the I-system would be involved within the semantic stream while the syntactic stream would a separate system for grammatical processing, as we can see in 2.6. One novel element coming from Baggio (2018) is the E-system and its role in the coordination between meaning, linguistic unit, and the state of the world. I present in greater detail the coordination aspect of communication in Chapter 3 and Chapter 5, but for now, the vital thing to keep in mind is that a model of linguistic prediction should explain cloze results by resorting to mechanisms or processes that could be linked with one or more of these processing systems. The three systems are represented in Figure 2.6 (Baggio, 2018, adapted from p.186).

In Figure 2.6, the R-system and the I-system are represented in blue, the grammatical stream in green, and the E-system in magenta. According to Baggio (2018), the first phase of the interpretation is related to the two processing streams described by Jackendoff (2007). The grammar stream, see the upper blue arrow, generates a morpho-syntactic analysis ('S') every time a new word is presented in the input. In the semantic stream (the lower blue arrow), each word's meaning is activated along with related lexical-semantic types ('R'). In the second phase of processing, the I-system builds a parse tree using a compositional process ('c') and computes an interpretation model from the semantic analysis ('i'). Together,

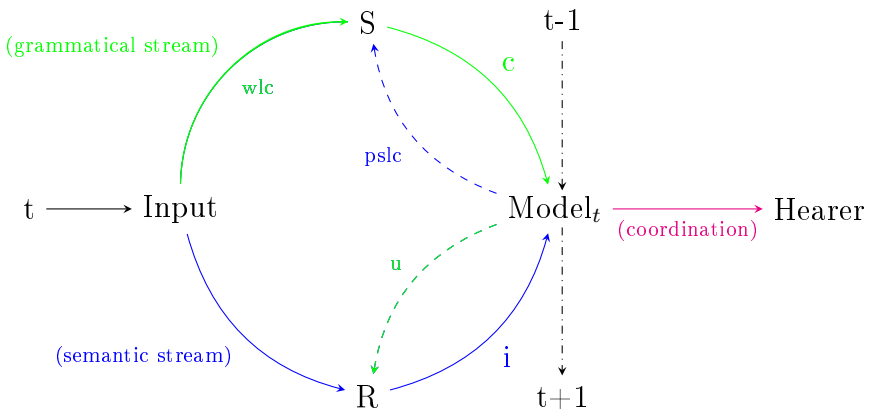


Figure 2.6 Cognitive architecture for Linguistic Processing

these two sub-processes build a mental model of the input.

It is interesting to note that the R-system and the I-system interact with each other at the morpho-syntactic analysis level ('S') and the level of the semantic analysis ('R'). In the latter case, the relational structures may be updated directly from the model when a new compositional contribution is processed. In the case of 'S,' it is formed both from the bottom-up constraints processed at the level of the word (wlc) but also from the update coming from the phrase and sentence-level constraints derived directly from the mental model processed by the semantic stream (pslc).[30] Finally, the last phase of this cycle is about the E-system and the coordination function between the mental model and the hearer's attribution of meaning. Additionally, it is essential to note that this coordination between the I-system's output and the E-system exists both from the interpretation and the production side of communication (Baggio, 2018).

The model of linguistic prediction I present in this thesis is a bit different from

[30]These phrase structure level constraints that act as predictive processes in input processing will be described in more detail in Chapter 5.

Baggio (2018) because it is about prediction and not interpretation. However, it is nonetheless meant to be compatible with the interpretative processing cycle presented in Figure 2.6.

2.5 Simulating a Linguistic Prediction

In this thesis, I am presenting a cognitively sound theoretically-driven model of linguistic prediction. The first thing to note is that I am not trying to simulate reading time (RT) or ERP results; instead, I am interested in generating predictions given a particular context because I think it might be a more direct way to assess linguistic prediction than ERP measures. After all, there is still much uncertainty regarding the ERP components and their relationship with specific linguistic information. For example, Bornkessel-Schlesewsky & Schlesewsky (2019) has described 12 different kinds of potential influences on the N400 component spanning from unexpectedness to lexical frequency to cross-linguistic differences. N400 readings alone, as helpful as they can be to help understand language processing, cannot directly answer the pressing questions regarding when and how anticipation is used in linguistic processing.

Furthermore, as was carefully argued by Nieuwland (2019), N400 results coming from isolated studies are challenging to reduplicate, which has the unfortunate consequence of hindering the firm conclusions drawn about prediction and language processing from some of the studies that are often cited as a reference.[31]

---

[31]Some have proposed to use measures of what is happening during language processing, i.e., the activation related to a prediction, instead of a measure of the result of this prediction like the N400 component (Grisoni et al., 2017; Aurnhammer & Frank, 2019b).

Some models of linguistic prediction are interested in modeling linguistic prediction in terms of N400 components. However, I do not discuss them in this thesis.[32] Instead, rather than modeling cognition, we are modeling the result of the cloze directly, i.e., the prediction of an upcoming word. This will help disambiguate between the respective roles of similarity, predictability, and plausibility within language processing.

[32]One of these models is the one proposed by Bornkessel-Schlesewsky & Schlesewsky (2019); Bornkessel-Schlesewsky et al. (2015) in which they present a predictive coding architecture to explain empirical results about the amplitude of the N400 component. Another model of linguistic prediction is the one developed by Rabovsky (2019, 2020); Rabovsky & McClelland (2020) which I briefly present in Chapter 6.

CHAPTER III

LINGUISTICS AND PREDICTION

After empirically circumscribing the cognitive structure of linguistic prediction, we have to make sense of this from the perspective of linguistics. In this chapter, I present a desideratum for linguistic tools to be compatible with, and then I present different linguistic approaches that could correspond to these requirements.

To better differentiate all the necessary components of linguistic prediction, I separate the treatment of linguistic information into two kinds of contribution: meaning compositions and coordination processes.[33]

Compositional processes follow from The Principle of Compositionality which tells us that the meaning of a sentence is a function of its parts and the way they are combined (Partee, 1984; Pelletier, 1994; Dowty, 2007). Compositionality is a process that combines words and smaller linguistic units into larger bits of information such as constituents and sentences. In terms of linguistic prediction, meaning composition is responsible for constraining the prediction using information present at various levels of granularity.

---

[33]This distinction is somewhat parallel to what Clark (1996) described as the difference between the Chomskian tradition that is focussing on the products of language, in my case, the product of composition, and the perspective from the pragmatics tradition that is focussing on the action of language, which I translated as coordination.

The other important aspect we need to consider is that communication is driven by consideration of coordination between speakers and hearers. This coordination aspect of linguistics includes being able to derive what our interlocutor is trying to use pragmatic clues and retrieving the correct meaning of a word. This view about coordination stems from the fact that "meanings are equilibrium points in a system of exchanges" (Gärdenfors, 2014, p.73) and also from the regularity nature of language (Gastaldi, 2020). This coordination is present at different levels of communication, and it usually involves some attribution of intentions.[34] Coordination processes are constraining the linguistic prediction using information that is not coming directly from the words expressed in the sentence. For example, the situation model from Kuperberg et al. (2020) that we described in Chapter 2 could be considered a result of this coordination between the two communicators. Once derived, the situation model constrains or limits the realm of possible continuations for a given truncated sentence.[35]

3.1 Desiderata for a theory of linguistic prediction

In order to use linguistic theory to model linguistic prediction, we need to write down a desideratum of features that the linguistic tools must be able to take into

[34]For example, Reboul (2011) makes a distinction between proximate and ultimate intentions because these are linked to different levels of communication. A proximate intention has to do with the speaker's informative and communicative intentions. In contrast, an ultimate intention is related to the speaker's behavior to induce the hearer.

[35]It is important to mention that the distinction between meaning composition and coordination processes is not naturally matched with the different disciplines of linguistics as semantics itself could be thought to appeal to both.

account. The first requirement is that of empirical adequacy.[36] In addition, we must also care about explanation, and we believe the most interesting linguistic theories are closely connected to cognitive realities. The following is a list of features that a theory of linguistic prediction should ideally meet. The first feature is *incrementality* because linguistic interpretation is also incremental, as was presented in Chapter 1. In addition, two critical features related to incrementality are *non-monotonicity* and *interpretability at a sub-compositional level.*

Before discussing each of these features in more detail, it is essential to note these features must hold for the two kinds of linguistic contributions since they are features linked with the structure of the linguistic theory, and this structure is independent of the nature of the linguistic meaning itself, be it derived from composition or coordination. Finally, these desiderata are meant to be requirements externally imposed on any processing theory of linguistic prediction.

3.1.1 Incrementality

Incrementality is the property of a system that processes its input one by one as soon as new input is encountered. For language processing, an incremental system processes every new utterance word-by-word incrementally so that the system can compute some output at every step along the way. At each of these steps, the output is a partial interpretation of the meaning of this incomplete utterance. The incrementality of a system has been described as a 'greedy' process because outputs are updated as soon as new inputs are processed (Michalon & Baggio,

[36] In this thesis, the empirical data comes from the results of Bloom & Fischler (1980) which will be discussed in Chapter 4 and Chapter 5.

2019), and the interpretation process does not wait for a complete utterance before debuting. In this thesis, I am interested in linguistic prediction at the word level, and I thus omit to discuss incrementality at lower levels of granularity. However, by requiring a linguistic model to process input incrementally at the word level, this should also reverberate at the level of the morphemes and the phonemes.[37]

Incremental processing view has been discussed for language production (Ferreira & Swets, 2002; Levelt, 2012), and it has also been discussed with respect to interpretation in psycholinguistics and computational linguistics (Ambati et al., 2016; Smith & Levy, 2013).[38]

As we saw in Chapter 1, if someone is asked to predict the next word of the utterance in (44), this person would be able to do it by interpreting the first part of the truncated sentence.

(44) John went to the park to fly a...

This person would also be able to predict the next word if the utterance was only "John went to the park to" or even "John went to the." When the utterance gets shorter, it becomes more difficult to guess the upcoming word because the prediction is less constrained, and we have more options for the words to choose from. However, it would still be possible to predict an upcoming word for any degree of truncation of a sentence. Incrementality is thus a compulsory feature of linguistic interpretation, and it is also by ricochet a requirement for modeling

---

[37]This lower-level incrementality has been observed at the syntactic and phonological level (Marslen-Wilson, 1973; Marslen-Wilson & Tyler, 1980; Tanenhaus et al., 1995)

[38]Incrementality fits well with the idea that a speaker does not plan all of his upcoming sentences at once, but instead "interleave planning and execution processes to maximize fluency as well as allocation of resources" (Ferreira & Lowder, 2016, p.235).

linguistic prediction.[39]

3.1.2 Interpretability of sub-propositional content

Naturally deriving from the incrementality of linguistic processing is the idea that sub-propositional content is readily interpretable by the hearer. Sub-propositionality comes as a consequence of incrementality because a partial interpretation must be updated every time a new input is processed. Propositions have been the basic unit of propositional logic and formal approaches of semantics for a long time (Moore, 1953; Russell, 1903, 1910; Stalnaker, 1976; Frege, 1884, 1984) and they have been linked with the evaluation of truth value. However, the incrementality forces us to process or evaluate content that is not yet truth-evaluable because not fully propositional. Therefore, the interpretative process must not use propositions or sentences as its basic unit and must be able to build integrative constructive inferences with heterogeneous information as soon as they can (Baggio et al., 2019, p.761).

(45) John went to the park to fly a...

Going back to our previous example, if (45) is processed incrementally, then every time a new word is heard, the system can use this new word to update the partial interpretation of the incomplete utterance. Starting from "John" then "John went," then "John went to," to "John went to the park to fly" the hearer is able, at

[39]Incrementality will be discussed in more detail in Chapter 6, but even though is a desideratum when developing a model of linguistic prediction, it will unfortunately not be part of the vanilla model presented in the next two chapters.

every step, to interpret every sub-proposition. Even if "John went to the" is not propositional, the hearer can still grasp its probable meaning and then interpret it. Interpretability of non-propositional content is also a requirement for a predictive theory because if we want to predict the upcoming word, we must also interpret non-propositional content.

3.1.3 Non-monotonicity

Non-monotonicity is linked with the interpretability of sub-propositional content. Non-monotonicity is a property of the consequence relation of logic when an inference is derived defeasibly. A *defeasible inference* is an inference that can be retracted when more information is taken into account (Strasser & Antonelli, 2019). Simply put, a non-monotonic inference is an inference that is not monotonic, where *monotonicity* is defined as in (46).

(46) If $\Omega \vdash \sigma$, then $\Omega \cup \Gamma \vdash \sigma$

The idea of non-monotonicity has already been discussed when it comes to discourse interpretation (Baggio et al., 2008, 2019). In discourse, the logic behind the derivation of inferences must also be non-monotonic since the inferences are defeasible. In other words, an interpretation derived from an utterance might change when new information is added, as in the following example taken from (Baggio et al., 2019).

(47) I began the novel in November. The pupils listened attentively. However,

some found dictation extremely boring.

From the first sentence, we may interpret that the novel is being written. The second sentence then suggests that reading it is the most likely action, and the third sentence turns the current interpretation around again. In this example, the interpretation of the first sentence changes every time a new sentence is processed.

What is true here at the discourse level is also true at the utterance level itself. For example, in a garden-path sentence like (48), the sentence structure's first interpretation is updated when the sentence is completed. When the first part of the sentence is processed, i.e., "The horses raced", *raced* is interpreted as a verb, but when the whole sentence is processed, then *raced* become a passive participle. Thus, the sentence has an interpretation, but it has to be re-analyzed or recomputed after the word *past* is processed because the structure assignment breaks down (Fodor & Ferreira, 1998).

(48) The horses raced past the barn fell.

Another argument for non-monotonicity is illustrated by the fact that a person can usually build a partial representation of an event even before attributing all the thematic roles for this event. For example, we could derive the event-structure *spreading something on bread* in (49) without having encountered the 'something' place holder yet. Deriving this event structure requires being able to interpret or process sub-propositional content, and here the output of the interpretative process is not a proposition but the structure of a proposition (Kuperberg, 2016).

(49) He spread the warm bread with ...

If we need non-monotonicity to render a dynamic interpretative process, it is also a condition we must meet when considering predictive processing.

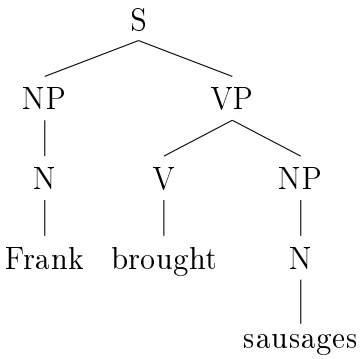
3.2 Language is about meaning composition

The Principle of Compositionality tells us that the meaning of a complex expression is derived from the meaning of its parts and the way they are combined. In linguistics, this usually implies that a sentence is composed of the meaning of words and from their syntactic combination. For example, in predicate semantics, we can translate the meaning of the sentence *Malika is strong and brilliant* into (50), where the meaning of the sentence is composed of two predicates that are combined using a conjunction sign. Additionally, these predicates are attributed to 'Malika' as an argument.

(50) $\text{strong(m)} \wedge \text{brilliant(m)}$

If we know the meaning of *strong(m)* and *brilliant(m)*, and the nature of a conjunction, we can derive the meaning of any sentences composed by these three elements. Composition also plays a role in the formation of syntactical constituents like the one in (51) because when we change the syntactical relationship between the words, it leads to a change in meaning (*Sausages brought Frank.*).

(51)



This section discusses syntactical and semantics approaches and presents how these two linguistics disciplines can deal with the desideratum introduced previously.

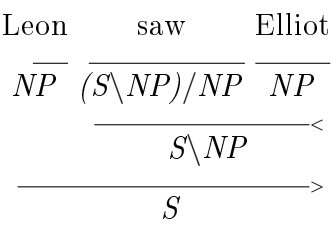
3.2.1 Syntax

Incremental theories have many uses when it comes to language modeling (Ambati et al., 2016; Morrill, 2000), and there is almost a universal agreement that incremental parsing is cognitively plausible (Marslen-Wilson, 1973; Sedivy et al., 1999; Milward, 1995). However, surprisingly, not that many approaches in grammatical formalisms are trying to integrate incremental features (Purver, 2015). There are many different theories of grammar (Steedman & Baldridge, 2011), and I do not plan to cover all of them. Instead, I focus solely on a kind of lexicalized grammar called Combinatory Categorial Grammar or CCG (Steedman, 1999; Steedman & Baldridge, 2011; Steedman, 2019) because CCG is better positioned than most syntactic theories when it comes to incrementality and non-monotonicity (Phillips, 2003).

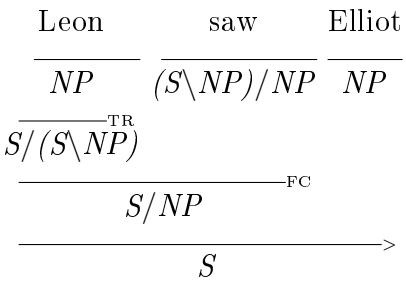
Combinatory Categorial Grammar (Steedman, 1996, 1999) is a context-sensitive

grammar formalism (Stanojević & Steedman, 2019) in which the linear order of the constituents and their interpretation are entirely defined by the lexical entries for the words (Steedman, 2014). In CCG, every word is assigned a syntactic type and the words are then combined using different combinatory rules that are entirely conditioned on those syntactic types (Ambati et al., 2015). Those combinatory rules are independent of the structure or the derivation (Steedman, 2014). In CCG, the notion of constituent structure is very flexible which allows for a natural integration of incrementality when constructing derivation trees for a sentence. This incrementality present in CCG allows for a flexible constituency since a constituent formed at the beginning of a sentence might not exist anymore when the full sentence is parsed, i.e. it is possible to have constituency conflicts throughout the incremental parsing (Phillips, 2003). In (52) we see an example of an incremental parsing using CCG:

(52)



(53)



CCG offers the possibility to parse the same sentence two different ways, either incrementally or not. For example in (53), by using what is called a Type Raising (TR) operation (Phillips, 2003) it is possible to combine *Leon* and *saw* together

first and then combine the output with *Elliot* by Forward Composition (FC). In non-incremental parsing, the semantic connection between *Leon* and the verb *saw* would only be established when the full sentence had been processed. In CCG, the meaning of "Leon saw" would be readily available as soon as it is processed (Stanojević & Steedman, 2019).

One consequence of this incremental construction is the constituent's flexibility, i.e., constituents may be built and destroyed during the incremental parsing (Phillips, 2003). This notion of constituent flexibility is crucial because it relates to the previous section's non-monotonic condition. Here the constituency is non-monotonic when building the structure of the sentence, and this a feature that is not present in dependency grammar where c-command relations are added monotonically (Phillips, 2003). Finally, CCG can represent a partial structure or partial constituent without any difficulty (Stanojević & Steedman, 2019). Thus, it would not be a problem to interpret partially derived sub-propositional bits of sentences using CCG and then combining them incrementally to form the complete sentence.

In summary, CCG checked all the requirements for a theory to be able to model linguistic predictions: it is incremental, it allows interpretation of partial constituents (which was labeled sub-propositional interpretability in the previous section), and its constituent flexibility opens the door for non-monotonicity.[40] I present this model because I think it could be a helpful framework moving forward for integrating the syntactic contribution for linguistic prediction. However, in the model presented in this thesis, the contribution from syntax was instead computed by hand for the reason of space and time.

---

[40]For this thesis, I do not discuss other syntactic theories that could also meet the requirements. For example, Dynamic Syntax (DS) (Kempson et al., 2001; Purver & Kempson, 2004) is also inherently incremental, but because of its monotonicity, it might be less natural than CCG to use it in a model about linguistic prediction.

3.2.2 Semantics

I do not discuss here the multitude semantic theories that could be used to build a theory of linguistic prediction. Historically, some semantic theory were not incremental by nature, but after the dynamic turn in semantics (Kamp, 1981; Heim, 1982), some semantic theory invented new ways to take into account incrementality in natural language semantics.

In Dynamic Semantics, time is taken into account, interpretation becomes a process, and the meaning of an expression or a discourse is incrementally updated every time a new bit of information is encountered. This view is in line with empirical results from psycholinguistics that support the idea of an incremental process for semantic interpretation (Sedivy et al., 1999; Tanenhaus et al., 1995). This perspective somewhat clashed with the static view of semantics put forward by standard predicate logic for which composition is not a process and for which interpretation must wait for the composition to produce a complete proposition before it may start (Baggio, 2018). On the other hand, in Dynamic Semantics, the meaning is seen as a potential of change coming from the context (Kamp, 1981; Heim, 1982). Within the Dynamic Semantic perspective, I choose to focus on the approach called Discourse Representation Theory or DRT (Kamp, 1981, 1995, 2008; Kamp et al., 2011).

In DRT, meanings are called Discourse Representation Structures (DRSs), a mental representation built by the interpreter. Every time a piece of new information is processed, the DRS is updated accordingly. The DRS content may consist of either an object referred to within the discourse or consists of conditions over which the objects are referred to. On their own, the DRSs are static, but their constant update makes it possible to represent the dynamic incremental process of

Table 3.1 Comparison between standard predicate logic formalisms and DRT

Standard predicate logic		
$\exists x[dog(x) \wedge bark(x)]$	$\rightarrow$	$\exists x[dog(x) \wedge bark(x)] \wedge black(x)$
DRT		
$x : dog(x), bark(x)$	$\rightarrow$	$x, y : dog(x), bark(x), black(y)$ $x, y : x = y, dog(x), bark(x), black(y)$ $x : dog(x), bark(x), black(x)$

semantic interpretation. The way the DRT is defined implies that it can represent semantic interpretation at both the sentence level and the discourse level as long as a new object or a new condition is introduced by the speaker. The difference between standard predicate logic and DRT is illustrated in Table 3.1 using the example (54) (Baggio, 2018, p.106):

(54) A dog barks. It is black.

In standard predicate logic, the occurrence of x in *bark'(x)* is not bound by the existential quantifier $\exists$, and this is problematic because we have to find a way to insert this anaphoric meaning within the scope of the quantifier (Baggio, 2018). As for the DRT framework, a new sentence causes the DRS to be updated with the new information. Similarly, when the new information *bark'(y)* is processed, it is added to a new DRS, and then the two DRSs are merged. This merge operation combines the two DRSs into one DRS containing all the referents and conditions of both. From there, it is possible to resolve anaphoric dependencies, and we can combine the two variables x and y because they both have the same referent (Venhuizen et al., 2018). This operation is possible because, in DRT, the

discourse referents are not bound variables under syntactic constraints coming from the quantifiers (Baggio, 2018).

Incrementality is naturally taken into account in DRT because it is a discourse processing account of semantic interpretation, i.e., a DRS is to be incrementally updated each time a piece of new information is processed. That being said, it is implied that the DRS is used to represent propositional content (Kamp et al., 2011), and it is never really discussed in the literature about DRT if it would also support lower levels of incrementality, i.e., word-by-word semantic interpretation.[41] This lack of consideration for a lower level of incrementality is also reflected in the lack of information about the capacity of DRT to integrate sub-propositional content because both features are linked with one another.

As for the non-monotonicity, according to the way the original version of the DRT was presented (Kamp, 1995), the merging between two DRSs always occurred between a DRS derived from new information against a DRS already containing the information presented in the existing context. As argued by van Eijck (2001), this causes a problem because it implies that DRSs merging is non-monotonic. However, what van Eijck (2001) sees as a problem, I see as a liability because non-monotonicity is required if we want to develop a model of linguistic prediction.[42] At the discourse level, monotonicity is sometimes called persistence or accommodation. Suppose there is no new information from the discourse that negates or terminates a specific referent or condition within the DRS. In that case, this state of affairs is deemed to persist to exist until such terminating information

---

[41] As discussed by Baggio (2018, p.115), this moderate version of incrementality is not the result of intrinsic limitations of the formal apparatus, but the lack of need to push incrementality at the word level to capture what is really of interest for DRT, namely anaphora resolution.

[42] Many updated models have since tried to resolve this issue by proposing different monotonic mechanisms (van Eijck & Kamp, 1997, 2011). However, those are outside the scope of this thesis.

arises (Kamp et al., 2011).[43]

3.3 Language is about Coordination

Communication is a communion of minds, and coordination is necessary for the communicative exchange to be successful and for communication to exist in the first place. If two people want to exchange information, then they need to use the same language, the same points of reference, and, on a certain level, a similar view about their knowledge about the world. For communication to be successful, it is a matter of linguistic units and a matter of the hearer understanding the intention of the speaker (Levinson, 2016; Hasson et al., 2018).[44] Some even argue that meaning emerges from this coordination between the world and the speaker's mental schemes and the hearer (Gärdenfors, 2014). To illustrate this need for coordination between the speaker and the hearer, we can turn to the example in (55).

(55) [Speaker A:] Take the blue square.

If we only have the utterance (55), it is not easy to understand what Speaker A meant, but if we add a little bit of context, it becomes relatively easy to retrieve the intended meaning. For example, if Speaker A had to choose between three

[43]DRT is not the only incremental approaches of dynamic semantics, and other frameworks could be also be envisioned. For example, File-change Semantics that was developed in parallel to DRT by Irene Heim (Heim, 1982, 2008) is also very interesting, although its monotonicity could potentially lead to complications when modeling linguistic prediction.

[44]Coordination between two parties is not the same as cooperation. The latter implies the former, but not the other way around.

squares (two red and one very pale blue), you would quickly conclude that speaker A wants you to take a very pale blue square. In another setting, if the three shapes were all blue (royal blue, very pale blue, and a greenish-blue), then you might interpret his 'blue' to mean something like 'stereotypical blue.' In these situations, the broader context influences the way the hearer interprets the words. In this particular case, it is as if the context directly influences the meaning of the word 'blue.' This example clearly illustrates the importance of the coordination aspect of communication for retrieving the information that the speaker intended.[45]

Gärdenfors (2014) distinguishes between two types of coordination: coordination about the meaning of a term, like in (55), and coordination of the knowledge about the state of the world (Gärdenfors, 2014, p.97). Coordination of knowledge is illustrated in (56), where we asked a participant to complete the sentence.

(56) John plays ...

Without any other clues, it would be difficult to guess the correct word. However, suppose I tell you that John is an avid chess player. In that case, it increases the probability that the next word be 'chess.' If I also mention that we are at a chess tournament and that the truncated sentence is uttered during the tournament, this would tip the scale for you to predict that the next word will most probably be 'chess.' In this case, the linguistic prediction is influenced by what information you have about the contextual setting of the utterance, i.e., what is your knowledge about the state of the world when this sentence is uttered.

---

[45]Determining the meaning of words from the interaction between the sentence and the word itself is described as wholism by Pelletier (2012) and is in agreement with this idea that words are ad hoc representations (Clark, 1996; Casasanto & Lupyan, 2015). Under this view, the meaning of a word is constructed from its use within a higher representational level, like the sentence level.

These two kinds of coordination are examples of mappings that need to happen between the world, the minds of both the speaker and the hearer. These mappings allow for reconstructing the mental representations of a speaker using the linguistic expressions that she used her.[46] This is very similar to what was described by Baggio (2018) when discussing the I-system of relations and the retrieval of referents. With a different mapping or different coordination, we could end up with different referents for the same expression. To avoid any potential misattribution of expression and referents, the hearer must make sure that his mapping function follows the speaker's mapping closely.

One thing that significantly helps the hearer to achieve just that is the regular nature of language. Language is about regularities at every scale, and this property opens the door to using mapping functions based on frequency or similarities (Gastaldi, 2020). I discuss similarity measures in Chapter 4, but for the moment, let me mention that the view described here is compatible with the words-as-cues views from Lupyan & Lewis (2017).[47] Under this perspective, a word is like a cue to access the inter-individual correspondence of meaning: mapping the world or the broader context and the expression. According to this view, the interlocutor himself is being taken into account, but this broader context can also include information about his emotions, preferences, moods, and motives (Van Berkum & Nieuwland, 2019). To understand how to model linguistic prediction, we must first underline the difference between the mapping between the words and the concepts and the one that is interesting for us, the mapping between the context and the concepts.

---

[46]Poeppel (2012) distinguishes between the *maps problem* which related with the spatial information about the activity in the brain and the *mapping problem* which has to do with the formal transposition of the representations that are dealt with in linguistics and neurosciences.

[47]Lupyan & Lewis (2017) rejected the word-as-mappings view where a word is a mapping between words and pre-existing concepts.

3.3.1 Coordination and Prediction

This section discusses linguistic prediction by comparing it with linguistic production and linguistic interpretation. To understand the nature of what a prediction is and what it takes cognitively to anticipate a linguistic unit, we also have to look at the production and interpretation of a linguistic unit. My goal here is to illustrate how production, interpretation, and anticipation are linked and present how anticipation would fit within the coordination aspect of communication.

3.3.1.1 Production

We usually say want we have in mind, and we use the production of an utterance to represent what we want to express. By producing an utterance, we change the world around us by updating it with some new information. Generally speaking, the directionality of the action is from the mind of oneself to the world. My purpose here is not to discuss all the intermediate steps involved in the actual production of an utterance but to differentiate it from prediction and interpretation.[48]

An utterance is a channel or a vehicle by which communication between the mind and the world is made possible. Of all the possible utterances you might use, you have to choose the one that best conveys what you think and, most importantly, the one that would be interpreted the way you intended to, given the particular

[48]For more information about the production mechanism, see Levelt (2012). In addition, recent studies have been investigating the intricacies between meaning production and phonological and grammatical articulation (Pickering & Garrod, 2013a,b).

context you are in. For example, if you want to convey that you are unhappy about something, you can always say something like (57), but you might also choose a more assertive approach like (58) or (59).

(57) I am not happy about this.

(58) I am angry about this.

(59) I could not disagree more about this.

The sentiment expressed might seem a bit stronger in the latter cases, but the thought that triggered these utterances could be the same in all three cases. So the same thought can be expressed differently, using different utterances.

To illustrate the importance of considering the broader context when choosing the utterance that best expresses our intended meaning, we can consider a situation where two co-workers, Ronald and Margaret, know each other but are not close in the office. Suppose that Ronald's parents are getting divorced, but that information was not shared directly with Margaret. After a couple of weeks, the news about the divorce was disseminated to almost everybody in the office, and thus Ronald knows that Margaret is aware of his parents' divorce even though he did not directly relay that information to her. While discussing at the coffee machine about Christmas and fruit cakes, Ronald tells Margaret:

(60) [Ronald:] My mother used to make one every year, but she did not make one this year.

(61) [Margaret replies with an acquiescing nod while pursing her lips.]

From Margaret's reaction, Ronald instantly assumes that she pursed her lips because she thought his mother's reason for not making a cake this year was related to the divorce. However, Ronald knew the real reason had to do with his mother not finding the ingredients to make the cake this year. Thus, Margaret misinterpreted that the reason was related to his mother's divorce, while Ronald intended no such interpretation. This example shows that even a simple utterance like (60) might convey additional non-intended meaning in some particular context.

The interpretation of an utterance has to do with its relationship between thoughts and the world, which is linked to the concept of 'Sense' as discussed by Frege (1892, 1984). According to Frege, we refer to the same thing using different paths of meaning. In this scenario, the thoughts are the referent, and the utterance is the sense that refers to this referent. This association between utterance and senses (or meaning) is already known to the speaker because, in his life, the speaker has already been gathering statistics of correspondence between utterances and states of the world. This knowledge allows the speaker to choose the corresponding utterance using a mapping function between utterance and world states. To choose an utterance to express oneself, one needs to find the best vessel to convey his intentions/thoughts from his minds to the real world, and this vessel is chosen using experience about corresponding states of the world and ways to express them. We can summarize this with the formula in (62), where f_1 and f_2 represent mapping functions between two different kinds of representation.

(62) $$\text{Utterance}_{\text{Prod}} \propto \Big(f_1(\text{thoughts,utterance}) \times f_2(\text{utterance,world}) \Big)$$

Utterance production is thus a function of two correspondences: the correspondence between thoughts and utterances and the correspondence between utterances and the world.

3.3.1.2 Interpretation

Even if it has some parallel with language production, language interpretation builds on a different relationship between language and the world. Now, the spoken utterance is an object of knowledge that the speaker has already created, and in order to interpret it, the hearer has to build a correspondence between an utterance and a state of the world. The idea is to retrieve the exact correspondence that the speaker used. Doing so, the hearer uses the same knowledge about coordinating world states and utterances as before, and she also considers the speaker's perspective.

Another aspect of interpretation is linked with the state of the world in itself. An utterance is like a hint to retrieve a new state of the world. Given this clue, the hearer can infer the new state of the world from the correspondence between utterances and states of the world, but the hearer also uses her knowledge of the world's current state to infer the updated state. Before producing the utterance, the speaker has some knowledge about what is true or not in the current world, and this knowledge allows him to infer new information and update the current state of the world. When a hearer encounters an utterance, she must combine this new information with the information she already possesses. The correspondence between world states and utterances has to be weighted according to the knowledge she already has about the state of the world at that time.

For example, suppose in a particular context, it is be deemed almost impossible for an athlete to break the world record in a given discipline at the Olympics. In that case, even if the hearer encounters an utterance that explicitly said that the world record was broken, the hearer has to take into account the unlikeliness of such an event for her to be able to interpret the utterance from the correspondence

between this utterance and the new world state it is describing.

The same correspondence between an utterance and a world state could lead to different interpretations in different contexts. Let me recall the previous example of Margaret and Ronald. This misunderstanding happened because the same utterance was meant in one way by the speaker. The hearer interpreted it differently because the latter did not use the same correspondence function as the former. The correspondence function is primordial if one wants to retrieve the intended meaning, but this correspondence might evolve according to the context and according to the different perspectives of the speaker and the hearer.

To sum up, the interpretation process is about combining this correspondence function with actual knowledge about the world and contextual probabilities about potential new world states. In other words, to interpret is to use what you already know about the world so you can find the correspondence between an utterance and what information it conveys, as in (63), where f_1 and f_2 represent mapping functions between two different kinds of representation. These indices are there to help differentiate the two mappings functions, and these should not be confused with the mapping functions previously expressed in (62). P is the probability distribution representing the state of the world.

(63)

$$\text{Utterance}_{\text{Inter}} \propto \big(f_1(\text{utterance,world}) \times \text{P}(\text{world}) \times f_2(\text{thoughts,world})\big)$$

Utterance interpretation is a function of the correspondence between utterances and the world, but it is also a function of the world itself. This second term en-

compasses the automatic update of world states from already acquired knowledge and the information about the probability of transiting from one state to another. This latter probability is related to the continuity of world states, i.e., if we are in state A, then we can reach state B by following a continuous path over different intermediate states that are all linked with each other.

In principle, interpretation would also involve epistemic judgements about said derived interpretation,and the last term in (63) is there to represent this relationship between the content expressed by the utterance and the judgement made by the hearer about this content. This relationship is sometimes called the *Reliability* of the speaker (McCready, 2014). If we also consider the believability of this content, we can more generally use the term *Epistemic Vigilance* mechanisms (Sperber et al., 2010) to refer to such processes.[49]

3.3.1.3 Prediction

If language production was about externalizing thoughts to the world, and interpretation was about retrieving information about the world, then a linguistic prediction is a combination of those two. A prediction is first generated via a production process which is also about the correspondence between the state of the world and utterances. Prediction is somewhat the reverse process because the goal is to anticipate a piece of information necessary for the hearer to retrieve the correct meaning expressed by an utterance.

For example, when a hearer encounters an incomplete utterance and then uses

[49]See Corbeil (2014) for a discussion about the role Epistemic Vigilance Mechanisms play during the interpretation process. However, this issue is beyond the scope of this thesis.

the correspondence between this incomplete piece of information and an updated state for the world, under some circumstances, this hearer might be able to use what she already knows about the world and this partial utterance to infer the missing piece of information. In other words, it is possible that, under the right circumstances, a hearer can guess the rest of the utterance "long before the speaker has completed its delivery" (Geurts, 2010, p.73).

When the hearer partially interprets an incomplete utterance, she can still update the state of the world, and then, from this updated state of the world, generate the missing pieces of information by using the same correspondence between utterances and states of the world that is used in language production. Under this view, a prediction is both an interpretation process and a production process: interpretation of the partially constituted utterance and production of the missing piece of information from the derived partial interpretation. The prediction generated by the hearer is the piece of information that, once combined with the incomplete utterance, would give rise to the same state of the world that the one which resulted from the partial interpretation process as in (64), where we again have two mapping functions f_1 and f_2, and a is the probability distribution P representing the state of the world.

(64)

$$\text{Word}_{\text{Pred}} \propto \big(f_1(\text{partial utterance,world}) \times P(\text{world})]_{\text{Inter}} \times [f_2(\text{utterance,world})]_{\text{Prod}}\big)$$

A prediction is the internal generation of an output O that would not change the state of the world if it was interpreted from the input. In our case, this output O is a word that would make the partial utterance complete. This prediction is

entirely internal, meaning that it is generated internally by the hearer, and it is not an external input (Michalon & Baggio, 2019).

A prediction is a generalization of an outcome; it is a process that goes beyond the actual data and creates new data (Phillips, 2018). Here, the interpretation of the partial information and the generation of a prediction are happening concurrently. In other words, the hearer starts attributing a corresponding updated state of the world (or conveyed meaning) to this utterance, and she then uses this partial interpretation to find a correspondence between this inferred new state of the world and the complete utterance that would have led her to this same conclusion about the updated states of the world. Finally, she can generate the missing information to bridge the gap between this updated state of the world and the incomplete utterance.

Under this perspective, a prediction is similar to an interpretation or a production, although the uncertainty associated with these processes is different. When it comes to production, the uncertainty is related to the representational process of matching an intention with a new state of the world and then matching it to a corresponding utterance. As for the interpretation process, the uncertainty is caused by the correspondence between the utterance and the world states and the uncertainty of the hearer's prior knowledge about this world state. In addition, there is an uncertainty in the correspondence between utterances and the world for both production and interpretation. When we consider linguistic prediction, a certain level of uncertainty is inevitable because the output of the prediction process is derived using both interpretation and production, and these processes have themselves uncertain outputs.

In this section, I have described predictions using utterances. However, I could also have discussed other linguistic levels. In this thesis, I am interested in predicting

upcoming words, but studies have already shown that other linguistic units may also be predicted in some situations.[50]

3.3.2 Pragmatics

Pragmatics is the discipline in linguistics that is interested in coordination. In this thesis, I espouse the view that pragmatics is about communicative actions and how they interact with conventionalized meanings (Gärdenfors, 2014). This notion of 'action' or 'use' has led pragmatics to conceptualize language as a 'game' played by two opponents or players (Lewis, 1969). These 'communicative games' are described and understood with some notions of strategy, intentions, and moves.

Over the years, many different pragmatics approaches have tried to understand how this coordination took place and what the consequences were for linguistic interpretation. The first kind of coordination about the meaning of a term led to the development of lexical pragmatics (Blutner, 1998; Wilson & Carston, 2007; Allott & Textor, 2012) while the other kind of coordination about the knowledge of the world fostered Gricean Pragmatics, Relevance Theory, and Game-theoretical Pragmatics, to name a few. For both kinds of coordination, a mapping is necessary to arrive at the fixpoints or equilibrium points that allow communication to be successful (Gärdenfors, 2014).

---

[50] See Chapter 2 and Huettig (2015); Nieuwland (2019); Levy (2008).

3.3.2.1 Gricean Pragmatics

Grice's work led to the development of a new perspective for the role of pragmatics in communicative studies and linguistics (Zufferey, 2016). The two principal contributions of Grice are related to his distinction between natural and non-natural meaning (Grice, 1957) and the advent of what is called conversational implicatures (Grice, 1989). In his view, the natural meaning is related to the external world, while non-natural meaning is instead linked with the linguistic meaning expressed in an utterance (Huang, 2016). Using this distinction, he developed a theory of communication and meaning in terms of a speaker's expressed intentions and a hearer recognizing those intentions. According to Grice, communication is based on communicative intentions, and it is viewed as resolutely inferential rather than purely conventional. In inferential communication, the speaker and the hearer use mind-reading abilities to retrieve the other's intentions. This kind of capability is required to understand communications like this:

(65) a. [Speaker A]: Are you coming to the field with us?
b. [Speaker B]: I have to go to the doctor at 4 pm.

In this example, Speaker B does not answer Speaker A's query directly, but we can infer that Speaker B will not go to the field with Speaker A. This inference about Speaker B not going to the field is made possible because of our belief that Speaker B must have had a reason to utter (65-b) instead of something else, i.e., if he had been going to the field, he would have said so more clearly. This inference is based on specific assumptions about the communicative behavior of a speaker.

Another example of this inferential view of communication is the 'pointing the

empty glass' gestures often observed at a restaurant or during a cocktail party. When you see someone across the room that points to the empty glass in her hand, you can infer that this person is trying to communicate something to you. The intention of this agent is for you to recognize that she means something and for you to infer that she probably wants you to bring her a glass full of what she was having before.

This agent intended this new meaning, and you were able to retrieve it using your knowledge about the world and your knowledge about what communication usually is. However, you could not associate this particular gesture to some sort of conventionalized meaning, i.e., you had to use a non-deductive inference to understand what she meant. This non-deductive inference is called abduction and is defined as an inference where the premises do not entirely warrant the conclusion. In other words, it is to find the best possible explanation for a given observation (Bunt, 2016). To illustrate this idea, we can use an example where the conventional meaning is pragmatically enriched:

(66) a. Fabio got the car to start.

b. Fabio got the car to start. In fact, he started it in the normal way.

In (66-a), the meaning explicitly expressed would be compatible with other situations where the car started, but (66-a) would often be understood as also conveying the information that it was difficult than expected for Fabio to start the car. This information is not entailed by the content of the utterance because it is defeasible, i.e., it is possible to cancel this information without any major problem like in (66-b). This notion of meaning has been introduced by Grice (Zufferey, 2016),

and it has been named a conversational implicature (Huang, 2016).[51]

3.3.2.1.1 Conversational Implicatures

Conversation Implicatures (CI) have two mean features or properties: defeasibility and non-detachability. We just illustrated the first one when we argued that a conversational implicature could be canceled by enriching the linguistic context, i.e., by adding another utterance as we did in (66-a). The other important property of CIs is related to them being retrievable from the meaning associated with an utterance rather than from the explicit content of this utterance. For example, if an utterance like (67-a) triggers an implicature like (67-c), then another utterance composed of different words, but conveying the same meaning as (67-b) could also trigger the same CI (Potts, 2004).

(67) a. Could you give me the salt, please?
b. Can you reach the salt?
c. [Conversational Implicature]: please pass me the salt.

The fact that an implicature is detachable from any specific item composing the utterance, but not-detachable from alternative ways of expressing the same meaning, distinguishes CI from conventional implicatures (Geurts, 2010; Zeevat, 2015a).[52] According to Grice, CIs are derived from some general principles related

---

[51]Grice also introduced the notion of *conventional implicature* although most of his writing is focussed on *conversational implicatures*. I will not discuss conventional implicatures here, but you can see Potts (2004) or Bach (1999) for a more thorough introduction.

[52]Grice (1989) also distinguished between two kinds of CI: Generalized and Particularized. However, this distinction is not deemed relevant by all scholars (Recanati, 2003; Russell, 2012; Carston & Powell, 2006; Geurts, 2010; Green, 1995) and I thus do not discuss it in this thesis.

to communication. These principles include a cooperative principle and a series of maxims (Grice, 1957). It is important to mention the maxims do not directly determine the CIs. Instead, the interpreter uses them as an implicit guide to retrieve these CIs (Bach, 2012).

The principles presented comes from the original approach from Grice (1989), but many adjustments or improvements have been proposed throughout the years. For example, Neo-Griceans approaches have proposed to reduce Grice's maxims to fewer principles (Huang, 2016): Horn (1984, 1995, 2010) put forward a Q-principle and R-principle, and Levinson (1983, 2000) a Q-principle, an I-principle and an M-principle. Others, like Relevance Theorists (Sperber & Wilson, 1995; Wilson, 2016; Carston, 2012), which are sometimes called post-Griceans, have reduced even more the number of maxims to keep only a modified version of the maxim of Relevance as we will see in the following subsection.

3.3.2.1.2 Prediction and Gricean Pragmatics

Grice defended the idea that communication was inferential and that Conversational Implicatures were retrieved with the help of maxims of communication. However, his approach was never meant to be cognitively realistic. Grice was an analytical philosopher and what he offered was a rational reconstruction of how CIs are derived. His proposals were not supposed to be viewed as psychologically sounded, but only as normative claims about the communicative behavior of rational individuals (Saul, 2002). Thus, it is difficult to discuss whether Gricean pragmatics would be compatible with the desiderata that need to be satisfied by a model of linguistic prediction because many different cognitive accounts could be descriptively compatible with the Gricean account. Despite this, it remains possible to discuss the three main features of predictions at a general level and

their potential correspondence within the Gricean account.

The first thing to note is that non-monotonicity seems readily achieved since a CI is defeasible when we add another utterance after the one that leads the interpreter to derive it in the first place. However, the Gricean derivation of an implicature is usually considered a post-compositional process. This means the implicature is derived only once the full proposition has been formed (Recanati, 2003). Bach (2012) calls the derivation of an implicature a post-propositional process. Under this perspective, conversational implicatures, as described by Grice, always 'remain external' to the explicit meaning of the utterance (Recanati, 2017, p.2). This implies that Gricean pragmatics would neither be incremental nor sub-propositionality interpretable because only when a complete proposition is formed would we be able to derive an associated implicature.

In terms of incrementality, a prediction about a conversational implicature could occur after the last word has been uttered, and this would be problematic for linguistic prediction. Similarly, an incomplete utterance would not be interpretable according to the classical Gricean view because only propositional content can be used to derived implicatures, not sub-propositional content. These two desiderata are not achieved because of the strict obedience to the notion of propositionality.

Even though nothing within the Gricean framework directly forbids us from treating an incomplete utterance as a complete one, relaxing the propositional constraint is not really in line with the spirit behind the Gricean account. Therefore, we should accept that the Gricean account seems to only accommodate one condition out of three, but this is not a big problem per se because other accounts in pragmatics can accommodate all three conditions.

3.3.2.2 Relevance Theory

Relevance Theory (RT) is often categorized as a post-Gricean theory because it develops Grice's ideas in line with psychology and cognition (Carston, 2012). Even if Grice might have inspired a psychological account of pragmatics, his central theme was mainly centered around philosophy and semantics, and his treatment of the derivation of an implicature was only rational reconstructions of how one mind might work (Wilson, 2016). Now, before discussing the compatibility of RT with the desiderata of a theory of linguistic prediction, I present a brief overview of some essential features of Relevance Theory.

3.3.2.2.1 Inferential approach to communication

If communication was purely about encoding and decoding, we should expect pragmatics to be neutral, i.e., communication should be symmetrical. This neutrality comes from the fact that the process of encoding the information is simply the reverse of the decoding process performed by the hearer (Moeschler, 2013). However, Relevance Theory rejects this purely encoding-decoding view and supports an inferential approach to communication.

> On this approach, pragmatic interpretation is ultimately an exercise in meta-psychology, in which the hearer infers the speaker's intended meaning from the evidence she has provided for this purpose. An utterance is, of course, a linguistically-coded piece of evidence, so that verbal comprehension involves an element of decoding. However, the decoded linguistic meaning is merely the starting point for an in-

> ferential process that results in the attribution of a speaker's meaning (Sperber & Wilson, 2002, p.2).

This framework acknowledges a crucial encoding-decoding component of communication and argues that some inferential processes are always involved in the interpretation process.

3.3.2.2.2 Relevance

Relevance is a double-sided property involving balancing the cognitive efforts it takes to process information and the cognitive effect that the processing will create. The more processing effort it takes to derive those effects, the less relevant it is; conversely, the more cognitive effect it has, the more relevant a piece of information is (Wilson & Sperber, 2012). The most important claim of Relevance Theory is that human cognition is geared towards optimizing relevance (Sperber & Wilson, 1995). In turn, this means that cognitive processes are influenced by the relevance of a piece of information and that the human mind only considers making an effort necessary to process something if it deems it relevant enough at that given moment. This constitutes the first principle of Relevance Theory (Wilson & Sperber, 2004):

- Cognitive Principle of Relevance
  Human cognition tends to be geared to the maximization of relevance.

The second principle of Relevance Theory is that, in communication, an ostensive stimulus will always be considered to be relevant enough to be worth processing it (Wilson & Sperber, 2004):

- Communicative Principle of Relevance

  Every ostensive stimulus conveys a presumption of its own optimal relevance.

This presumption of optimal relevance is primordial because it is the trigger that allows the interpretation process to begin. Even in cases where the satisfactory level of relevance seems not to be attained, the hearer will likely assume that the speaker has tried to be optimally relevant and failed to do so (Sperber & Wilson, 1995). The hearer will indeed begin the interpretative procedure because he thinks that the utterance is optimally relevant. He will first recognize the utterance as a communicative act and then accepts "the presumption of relevance it automatically conveys" (Wilson & Sperber, 2012, p.236) in order to begin the procedure to retrieve the intended meaning.

3.3.2.2.3 Interpretation procedure

The interpretation procedure follows a path of least effort until the expectation of relevance is fulfilled. This procedure can be broken down into two steps (Carston, 2002):

1. Follow a path of least effort in computing cognitive effects: Test interpretative hypotheses (disambiguation, reference resolutions, implicatures, etc.) in order of accessibility.

2. Stop when your expectations of relevance are satisfied.

Interpretative hypotheses are derived in order of the amount of cognitive effort necessary to process them (from least effort to most effort). The derivation continues until one hypothesis can reach the expectations of the hearer's relevance

at that moment. These expectations of relevance are defined as the threshold at which an interpretation is considered relevant enough for the addressee. The process will only stop when the expected cognitive effect is worth the cognitive effort to derive the interpretative hypothesis. Hence, in a case where the expected cognitive effect is wholly indeterminate, the hearer is likely to choose the most accessible interpretative hypothesis, i.e., the less cognitively demanding one, as he would consider it relevant enough for a lower cognitive effort (Sperber & Wilson, 1996).

3.3.2.2.4 Prediction and Relevance Theory

Now that we have introduced RT, we can describe whether it complies with the previously mentioned desiderata. The RT interpretative process is incremental, even if it is never explicitly described in the RT literature. However, the incrementality here is related more to the process itself and less with partial information. What is incremental in RT is the way an implicature is derived using different steps along the way. Whenever a new assumption is derived, or a new piece of information is used to infer something, it is a new step towards the complete interpretation of the utterance.

This idea is very similar to what we discussed for Grice because here, the process is incremental in terms of the rationality behind the interpretation and less in terms of interpreting partial representation right away. Although this other type of incrementality is not incompatible with RT in principle, it was never discussed explicitly in the literature.[53]

---

[53] An incremental view of Relevance Theory is described in Corbeil (2014, 2015). According to this view, the interpretation process of RT is incremental below the utterance level since every time new information is processed, be it coming from a word or coming from any other

With incrementality not being an issue in RT, we must also discuss sub-propositional interpretation. One of the central premises behind RT is that utterances are underdetermined (Carston, 2002), which means that an utterance by itself becomes propositional only when pragmatically enriched, is really in line with what was described earlier when discussing the interpretation of a partial sub-propositional sentence. In the case of RT, the notion of propositionality is only linked with the final form of the interpretation, namely the explicature and the implicature, but it does not play an active role in the interpretation process per se. In other words, propositional or sub-propositional, an utterance will be enriched and allow the derivation of an implicature from it.

Finally, it should not be surprising that non-monotonicity is also compatible with RT. As it was also discussed about Gricean pragmatics, implicature is non-monotonic and the procedure to pragmatically enrich an utterance to form a proposition is also non-monotonic because it remains defeasible. It is thus apparent that RT checks all of the boxes we need for a model of linguistic prediction.

3.3.2.3 Game Theoretic Pragmatics

Game theory has been used to understand interactive decision-making within different fields such as sociology, economics, and linguistics (Franke, 2013). It offers a mathematical way to study strategic interaction between multiple rational agents called 'players' where one decision's outcome is influenced by the decisions of others (Allott, 2006; Jäger, 2014; De Jaegher, 2008; van Rooy, 2004; Clark,

---

perceptual input, it automatically updates the state of the world by which the interpretation is derived from.

2012).

An easy example of such a strategic interaction is a situation where you have a choice to make when getting dressed for your prom. During the prom, a person wants to show her personality by wearing something bold, but at the same time, this person does not want to be considered to be too bold and also does not want to wear the same outfit as somebody else. The decision about which outfit to wear is strategic because the boldness has to be weighted with respect to the possibility that one made the same outfit decision. In this simple game, the outcome of this choice of outfit also depends on the choices made by other people playing the same game.

In this kind of game, the strategy is to weigh what you think others will do while keeping in mind that each player can have her strategy of action. Additionally, every player also has their *preferences* over the possible outcomes of the game (Jaeger, 2008). In this case, not everyone has the same tolerance for boldness or the same need for being recognized by a crowd of people, and everyone has their preference ordering for the outcome.

In every game, different outcomes lead to different utilities for the player. In game-theory, utility functions are related with preference ordering using the relation $\preceq$ (Binmore, 2009):

(68) $\quad u(a) \leq u(b) \quad$ if and only if $\quad a \preceq b$

This means that the utility associated with the outcome a is bigger than the one associated with the outcome b because a is preferred over b. A player will thus use a strategy that will enable her to maximize her utility (Benz et al., 2005). Usually, utilities are represented by real numbers attached to every possible

outcome, and these utilities are different for each player (Jaeger, 2008). Because most games involved coordinated action, like the outfit example, a player's action cannot determine by itself the outcome of a given game. Thus, the player has to consider the other players as well, and there are, in general, three families of games with respect to the kind of decision one may choose.

If a player already knows which of his actions will lead to specific outcomes, we say that this decision was taken under certainty (Benz et al., 2005). On the other hand, if the player can assign a probability for every possible decision, we can label this a decision under risk. Finally, if the player cannot assign any probabilities, i.e., if he does not even know the potential outcome of a given decision, he has taken his decision under uncertainty. Most papers on game theory focus on risky and uncertain games because real-world decisions under certainty are rare. An example of a game of uncertain decision is the classic Rock, Paper, Scissors game.

Game theory allows modeling the best solution for a particular game by looking at possible outcomes. Using utilities and the information that the other players also have access to the same board, we can predict how players will behave in given situations. The best solution will always be the one that is the most rational according to the agent. In other words, a player has to use a rational strategy to solve the game to his advantage, i.e., to maximize his utility.[54] In Rock, Papers, Scissors, the best strategy for both players is to play every move with equal probability.

Zero-sum games have primarily been used in logic and truth-conditional semantics (Clark, 2012). Communication is a bit different because both players can win if they participate in successful communication. When a hearer grasps the meaning

---

[54]A Nash equilibrium is a "strategy profile where each player believes the other players know his strategy in advance" (Jaeger, 2008, p.408).

expressed by a speaker, both players benefit from the interaction, and thus we cannot use a zero-sum game to represent this kind of game. Instead, when the player's utilities are aligned like in a conversation, we use a coordinating game or, more particularly, a Signalling game (Lewis, 1969, 1979).

3.3.2.3.1 Signalling Games

Language games have been first developed by Lewis (1979, 1969) and their treatment has been continuously updated and developed since then (van Rooy, 2004; Franke, 2009, 2010).

In this thesis, I am assuming that coordination between two agents or players is happening at the language level, meaning that when two people are discussing, they have to find a way to coordinate their linguistic strategy profile in order to communicate successfully (Jäger, 2014).[55]

Signaling games, as described by Lewis (1969), is composed of two players, a sender (a speaker) and receiver (a hearer) (Franke & Jäger, 2013). The main reason Lewis (1969) invented the Signalling game was to explain how conventions of meanings were maintained and used by interlocutors (van Rooij, 2004; Franke, 2016; Franke & Jäger, 2013). In a Signalling game, a receiver has to react to a signal produce by the sender, this signal being about private information that the receiver lacks (Franke & Jäger, 2013). In other words, the sender has privileged information about the state of the world, chooses a signal, and sends this message to the hearer, which will then act upon receiving this message (Franke, 2016).

---

[55]Another view was one of Prashant Parikh (2000); Parikh (2001) that looked at communicative situations as coordinated by rational behaviors of the agents without having to recourse to any specific linguistic constraints. However, I will not discuss this view here.

This action directly responds to this message, and this response can be physical (closing a window) or epistemic (believing in something). The Signalling game is successful when the response from the receiver matches the state of the world that the speaker observed. In a Signalling game, an utterance selected by the sender has no predetermined meaning, and the meaning emerges only from the strategic interaction between the two players (van Rooij, 2004).

A Signalling game thus involved two moves: one by the speaker and one by the hearer. The speaker uses a correspondence function from a state of the world to signals (or utterances), while the hearer uses a correspondence function from signals to states of the world. In a successful Signalling game, the strategy used by the hearer and the one used by the speaker will lead the hearer to find the best action in response to the signals sent by the speaker, and the outcome or the utilities of both players will depend on both correspondence functions (Franke et al., 2012). For example, if the set of possible states of the world is represented as S_W and one particular state as s_1, then:

(69) $s_1 \subset S_W$

(70) If $m \in M, \implies [\![m]\!] \subset S_W$

In (69) and (70), M is the set of all possible messages available to the sender, m is a particular message, and $[\![m]\!]$ the semantic meaning of m. The receiver has access to the set of actions A, and he picks one action a in response to the message m. In the simplest situations, the action available to the hearer is to update his beliefs to correspond to the state of the world, so we might say that $A = T$ in most cases.

From these, we could thus write down the correspondence function for the sender

and for the receiver (Franke et al., 2012):[56]

(71) Correspondence function$_{\text{Speaker}} = T \rightarrow M$

(72) Correspondence function$_{\text{Hearer}} = M \rightarrow T$

We can illustrate the importance of communication and Signalling games with an example adapted from Jäger (2014). Let us suppose that Margot and Fiona are planning to go to the movie theatre. When they arrived at the cinema, Fiona has to choose between two different kinds of movies, both starting at the same time: movie A and movie B. Fiona wants to pick a movie that Margot would also want to see, but they both do not know the other's preferences. This situation can be formalized as in Table 3.2 (Jäger, 2014, adapted from their Table 1).

Table 3.2 Example of Signalling Game

		Player 2 (Fiona)	
		a_1	a_2
Player 1 (Margot)	s_A	(1,1)	(0,0)
	s_B	(0,0)	(1,1)

In Table 3.2, we see two possible states of the world: s_A where Margot prefers movie A, and s_A where Margot prefers movie B. Fiona has a choice to make about which movie they will see; she could choose movie A (a_1) or movie B (a_2), but she has no idea what state of the world we are in. Then we can represent potential payoffs or outcomes for possible actions with respect to the state of the world. This table shows that choosing the wrong movie will benefit neither as

[56]Lewis (1969) assumes that communication is costless, but I return to the issue of processing cost in Chapter 6.

their payoffs would be 0 each while picking the right movie will give them a payoff of 1 each. The average payoff for both of them will thus be 0.5, but they can raise this average if they communicate because if Margot can inform Fiona of her preference, then Fiona will be able to choose a better outcome for the two of them. For example, if Margot tells Fiona "I like Movie A", Fiona will adapt her actions accordingly. Both players are better off when communication is involved, but communication also opens the door to some potential ambiguities.

If Margot says "I like Movie A" while in s_A and "I like Movie B" while in s_B, then Fiona interprets "I like Movie A" as meaning they are both in s_A, and interprets "I like Movie B" as meaning they are both in s_B. This mapping from states of the world to utterances can be written as a function F_1. Another option for Margot would be to say "I like Movie B" while in s_A and "I like Movie A" while in s_B. This utterance could lead Fiona to interpret "I like Movie A" to mean they are both in s_B and "I like Movie B" to mean they are both in s_A. This mapping function between states of the world and utterances could be called F_2.

One problem here is that Fiona has no way of deciding which mapping function Margot has used, which means that Margot could well be using F_1 while Fiona thinks that she used F_2. If there is a mismatch like this, then the two will have the worst payoff, i.e. (0,0). Generally, it would be more plausible that Margot used F_1 because we can assume Margot is a rational speaker that follows the Quality maxim from Grice. However, we generally must not rely on honesty or credibility to choose our path of actions (Jäger, 2014).[57]

To explain how to decide which mapping function a speaker is using, we have to use an iterative inference process conditioned over different levels of beliefs regarding

[57]Franke (2011) defines a particular class of Signalling games called an Interpretation game in which basic Gricean assumptions of cooperativity are implemented.

the speaker and the hearer. This back-and-forth process is about determining the beliefs the speaker and hearer respectively have about each other's beliefs (Franke & Jäger, 2013). This step-by-step strategic iterated reasoning has already been integrated within an approach that takes into account the meanings of the signals used by the speaker, and these models have been called Iterative Best-Response reasoning (Franke, 2011; Franke & Jäger, 2013; Franke, 2009; Franke & Jäger, 2013).[58] These iterated approaches are very interesting when modeling pragmatic processes, and I discuss probabilistic approaches in Chapter 5.

3.3.2.3.2 Evolutionary Game Theory

Evolutionary Game Theory (EGT) was first proposed by Smith & Price (1973) where they used it to study the Darwinian natural section. The basic idea behind EGT is that instead of single agents, the players are populations of individuals (van Rooy, 2004). When these populations interact, it affects these populations' reproductive rate, and the utility for each outcome represents the expected number of offspring of this population. The strategies are considered to be a genetically determined disposition of behaviors (Jaeger, 2008). Put differently, EGT is interested in determining the best strategies so that a population has the best reproductive rate possible.

We could illustrate this idea of evolutionary game theory by transposing it into our Rock, Paper, Scissors example. If you play this game once, you will probably have no idea how to choose. However, if you play the same game with the same opponents for 5, 10, or even 100 times, you will adapt your strategy along the

[58] Some other version of iterated game theory can be found in Jäger (2014) and Benz et al. (2006).

way: you will slowly modify your prior expectation of what the other player will do in light of the set of this past actions. The strategies you have will change when a series of the same game is considered (Benz et al., 2005).[59]

This kind of evolution is categorized as *vertical evolution* by Franke (2016) as opposed to what is called *horizontal evolution*. Vertical in the sense that populations of agents converge on a single mapping function between a signal and a state of the world. This *vertical evolution* is responsible for the conventionalization of the mapping function.[60] On the other hand, *horizontal evolution* is about learning more about the strategies employed by a specific agent throughout a series of games (Skyrms, 2010).

If we transpose EGT into our Signalling game, we get that the outcome is not the number of offspring, but the likelihood that a strategy or an action is imitated throughout many iterations (Benz et al., 2005) because strategies that lead to a higher utility are more likely to be played again after. Having a series of games will thus make some actions better than others in the long run, and it is those actions that the evolutionary process will naturally select. In linguistic terms, the mapping function we had between Margot's states of the world and utterances will slowly be conventionalized during the evolutionary process, leading Fiona to have a fixed strategy over time.

This stable evolutionary solution is called an Evolutionarily Stable Strategy (ESS)

---

[59]It is essential to note that being able to play a large amount of time also make a difference when playing a zero-sum game like Rock, Paper, Scissors or even when playing a coordination game like the famous prisoner's dilemma. See Clark (2012) for a thorough description of this classic game.

[60]The process of conventionalization is related to the fact that the more often a meaning is interpreted, "the more entrenched it becomes, resulting in easier access and faster processing" (Christiansen & Chater, 2016, p.18). A convention here broadly refers to a crystallized mapping between a concept and use and not the usual interpretation of a social code of conduct.

(van Rooy, 2004). EGT here can explain how conventionalization takes place between interlocutors during a series of conversations or dialogues. It is analog to what Dawkins called memes, i.e., cultural traits that spread from person to person by imitation is a cultural system (Dawkins, 2006). After a series of games between Fiona and Margot, Fiona will be better at guessing the probable mapping function that Margot uses because she will have experienced the different payoffs for her different guesses.

Slowly, the probability attached to either mapping function will fluctuate according to the retrieved payoff at every cycle, and after many games, one mapping function will become more probable than the other until one of them becomes almost inevitable. This fixation of the mapping function's probability will lead Fiona to choose a stable strategy of action. This conventionalization process is fundamental because, as we will see in Chapter 5, the mappings between the sentence-level representations and the contextual level representations play an important role in my predictive model.

3.3.2.3.3 Prediction and Game Theory

The core idea behind game theory is to model the behaviors of rational players to predict their strategy of actions. Concerning the incrementality feature, the kind of games I described in this section are static games like Rock, Papers, Scissors, or, when considering evolutionary game theory, a series of static games. Rock, Papers, Scissors is static because the players have only one action to perform simultaneously, not in the sense that this game does not involve any dynamic reasoning about the beliefs of the other player.

However, there is another extension of Signalling games where there are two speak-

ers or senders at the same time (Barrett, 2009), and this situation can be shown to be equivalent to one speaker sending two consecutive signals (Skyrms, 2010). In this consecutive game, also called a "syntactic game", a receiver gets a series of signals and has to perform an action for every one of these signals (Franke, 2016). A syntactic game involving a speaker sending a series of signals where a receiver acted every step would be incremental. It also seems that the interpretability of sub-propositional content would not be a problem for a syntactic game because it would always be possible to divide a composed message into smaller components and then treat them as different separate games.

The non-monotonic aspect of prediction is a bit more challenging to assess in classical game theory. However, it follows naturally from the iterated version of game theory, where an iterated inferential process occurs between the speaker and the hearer. When a hearer represents the speaker's beliefs, it corresponds to the first level of iterated response. The second level would be for the speaker to build a representation of the hearer's representation of the speaker's beliefs. This iterative process could continue until one of the players' cognitive resources become depleted or when one of the judges that the other one has already stopped this iterative process.

At every iterative cycle, there is a certain probability that the representation of the other player's beliefs is updated and changes a bit. In other words, the representation of the other player's beliefs is defeasible because, at every cycle of iteration, this belief could change and be adjusted. To illustrate this defeasibility, we could transpose this idea of iteration into the Gricean program. This move is not too far-fetched considering the natural connection between game theory and Gricean pragmatics (Stalnaker, 2006). In the Gricean approach, we most commonly assume that the first representation of an utterance is a literal interpretation, but when one of the maxims seems violated, we can revise the first

interpretation (Sperber & Wilson, 2012). One main advantage of iterative game-theoretic approaches of pragmatics is that it bypasses the need for having maxims in the first place (Franke & Jäger, 2013).

In this discussion about the desiderata of a predictive model of linguistic meaning and game theory, I omitted anything about evolutionary game theory because it does not play the same role as classical or iterated game theory. EGT is interesting because it explains why and how conventionalization arises over time, and it will play an important role when determining the correspondence mapping function between meanings and utterances. However, predictive behaviors are not to be modeled directly from EGT. The conventions derived from the evolutionary perspective are used to determine the best action to take, i.e., the best word to predict, but it is only indirectly considered in the iterated inferential process of prediction.

3.4 A Model of Linguistic Prediction: Interim Summary

In the first chapters of this thesis, I have presented empirical evidence describing the constraints imposed upon the derivation of a linguistic prediction during language processing. In addition, I have proposed desiderata for a theory of linguistic prediction. In this section, I summarize what we have discussed so far, and I introduce a summary of the coming chapters.

3.4.1 A Cognitive Model

The goal of the model presented in this thesis is to simulate how linguistic prediction is derived when participants complete a truncated sentence. In Chapter 2, we separated semantic and syntactic processing into two streams. When it comes to the derivation of a linguistic prediction, these two kinds of processing streams give rise to different types of constraints that contribute to the derivation of a prediction. The syntax constrains the probability value for a particular syntactic category type of the predicted word, while the semantic constraints are responsible for predicting the meaning that would best correspond to the rest of the sentence. Furthermore, in Chapter 2 and this chapter, we saw arguments that syntactic constraints are not usually sufficient to derive a linguistic prediction, and we presented the view that the semantic stream has precedence over the syntactic stream. The model of linguistic prediction that is developed in this thesis is thus centered around the semantic constraints that lead to the derivation of the meaning of a prediction.[61]

In the model developed in this thesis, the semantic prediction results from combining different predictions derived for different linguistic units at the lexical level and beyond. These different linguistic units are the output of the meaning composition process. When two words are combined (e.g., *red* and *car*), the composed meaning, in addition to the lexical meaning, also comprises the combination of semantic properties of these words and the relational properties that bind them. Once these predictions between the meaning expressed in the sentence and the predicted word are combined, we can compute a conditional probability of occur-

[61] The relative role of the syntactic constraints compared to the semantic constraints is discussed in Chapter 4.

rence corresponding to the predictability value we are looking to generate.

In addition to the contribution coming from meaning composition, we also have to consider the coordination aspect of a linguistic interaction responsible for the choice of mapping function that helps retrieve the meaning of a sentence within a particular context. In the model of linguistic prediction developed here, the contribution from the coordination aspect of language is twofold: it allows the hearer to derive a representation of the context, and it allows the hearer to use this contextual representation to constrain the derivation of a linguistic prediction. This contribution from the coordination aspect of linguistic interaction is discussed in Chapter 5. In Chapter 4, I present a language model that integrates the contribution from the meaning compositions.

3.4.2 A Language Model

A language model is a model designed to predict the next word of a sentence (Smith, 2019). It is usually written as the conditional probability of a linguistic representation given a specific context (Armeni et al., 2017). In this thesis, the linguistic representation that is predicted is a word.

(73) $P(\text{word}|\text{context})$

The conditional probability in (73) is also called the predictability of a word given a context. A language model is thus a model of predictability, and these probabilities can be derived from different approaches (Hale, 2016). In this thesis, I am interested in providing a principled explanation of the derivation of this

conditional probability.

Some approaches to cloze tasks have directly equated cloze scores with conditional probability and then argued that predictability was the sole influence when it comes to reading time results (Smith & Levy, 2011). However, some other models have proposed that semantic similarity also played an important role. In Chapter 2, we mentioned that Roland et al. (2012) argued that the effect of similarity should be taken into account independently from the effect of predictability. In other words, it is the combination of both the predictability and the similarity that influences the participant to choose a specific word.[62]

If we further develop this idea, we could argue that words predicted during a cloze task are the words that have been the most activated by different factors, be it from syntactic or semantic influences. This idea has been defended by Staub et al. (2015) when they argued that a cloze task is a task where multiple influences are used to activate different words. When a particular word reaches a threshold of activation, only then does it become a possible valid continuation. A similar idea about the general pre-activation across multiples domains of language processing is also defended by Brothers & Kuperberg (2021). Taking into account different activation levels is compatible with the idea that different words are processed with different levels of difficulty. In that case, low cloze words are less activated, and high cloze words are the words that are more strongly activated by the preceding linguistic context. Within this activation model, cloze score would be correlated to the relative level of activation (Staub et al., 2015).

Approaches based on surprisal are interesting because they give us a mathematical model of the effect of this conditional probability and the cognitive effort

[62]Their argument is compatible with surprisal theory, as it does not rule it out, but it states that there is an additional similarity effect that can top it out (Roland et al., 2012).

needed to process the new input. However, they do not inform us about how this probability is derived during the cognitive process because the notion of surprisal is independent of how this probability is derived. As described by Hale (2016), these expectation-based approaches are sitting at the computational level of Marr's analysis, and they are not directly linked with any psychological process per se. However, when developing a cognitive language model, we must make sure it is compatible with these computational-level theories.[63]

In following chapters, I present a language model based on the pre-activation of different levels of representations. This pre-activation is related to the similarity measures between different linguistic units, but it is different than what was described by Roland et al. (2012). In Roland et al. (2012), they argued that similarity measures were independent of the surprisal effect. I take a slightly different path in the language model presented here because I treat similarity as contributing indirectly to predictability via an activation-based semantic network. Since my focus is on developing a model of predictability, I leave to one side the question of how predictability is then linked to processing difficulty because it is outside the scope of this thesis.

[63] Conversely, surprisal theory is also agnostic with respect to the nature of the linguistic representations used to compute the conditional probability (Futrell et al., 2020).

CHAPTER IV

LINGUISTIC PREDICTION AND MEANING COMPOSITION

This chapter presents a language model that integrates the contribution from the meaning compositions. The backbones of this model are the Similarity Spaces (SS) and the word embeddings. Similarity spaces are used to measure the similarity between word embeddings. This value is directly linked with the derivation of a specific word's prediction given an incomplete sentence. In addition, I argue that meaning composition involves different levels of granularity associated with their compositional units. Finally, I detail the procedure by which we can derive the contribution from these different representational levels to retrieve a linguistic prediction, and I show some worked out examples. At the end of this chapter, I revisit the distinction we already discussed in Chapter 2 between similarity, predictability, and plausibility in terms of these representational levels.

4.1 Similarity Spaces and Word Embeddings

According to the distributional approaches of linguistics, the meaning of a word comes from its use; or as Firth (1957, p.11) puts it: "You shall know a word by the company it keeps". To retrieve the meaning of a word or at least to circumscribe

it, you can look at the other words surrounding it. Under this perspective, it should be possible to define the meaning of a word in terms of its neighbors within a particular sentence. When using a big enough corpus, every lexical entry is differentiated because every word is not precisely appearing in the same contextual windows, e.g., a couple of words before and a couple of words after the target word.[64]

Classically, this particular operation of extracting a word's distribution involved creating a co-occurrence matrix that would represent words as rows and contextual elements as columns (Li et al., 2004). This matrix is be derived by counting how many times a target word occurred in a particular context. For example, if we pick the word *candies* as the target, we can build a row matrix that represents the number of times another word is co-occurring with it in a linguistic corpus. Thus, this row-matrix or vector represents the meaning of the word *candies* in this corpus.

	sugar	children	sweet	dessert
candies	2	3	3	1

Figure 4.1 Example of the vectorial representation of the word *candies*

Word vectors can be derived manually by looking at the distribution of neighboring words, but it is also possible to derive them using machine learning mechanisms. Word vectors derived using the learning techniques in Natural Language Processing (NLP) are usually called *word embeddings*.[65]

---

[64]Distributional approaches typically use corpora having a vast number of words like the British National Corpus (10 million words) (Burnard, 2007) or ukWaC (100 million words) (Baroni et al., 2009) which allows choosing a smaller window without jeopardizing the precision-score (Kiela & Clark, 2014).

[65]The term *vector* is usually used in the Distributional Semantics literature, while the

A similarity space is a space where the relative position of different objects is determined by their similarity. For example, if we represent three word-vectors on the same 2d plot, we can compare two pairs of words based on their relative position in the similarity space. In the following example, we can readily see in Figure 4.2 that w_1 and w_2 sits closer to each other than w_1 and w_3. In this case, if we posit the euclidean distance to be a valid measure of similarity, then we can understand that w_1 and w_2 are more similar than the pair w_1 and w_3.

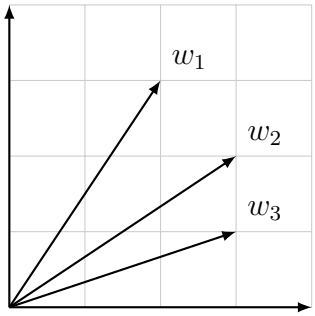


Figure 4.2 Representation of three word-vectors in a 2D similarity space

Word embeddings are a useful computational tool for linguistic modeling, but they are also interesting for modeling cognitive processes more generally (Mandera et al., 2017). For example, in cognitive science, vectors are also a natural way to model prototype theory (Rosch & Lloyd, 1978), but the main difference is that prototype vectors are directly derived from questioning participants, and they are not automatically derived from frequency measures (Turney & Pantel, 2010).

---

more general term *distributed representation* is mostly used in the literature about connectionism (Hinton et al., 1986; McClelland et al., 1986). Throughout this thesis, I use the term *word embeddings* which is widely used in computational linguistics to refer to such representations of meaning.

4.1.1 Deriving Word Embeddings

One limitation of the classical way of deriving word vectors is that, by counting the number of co-occurrences, we end up with huge sparse vectors, i.e., high dimensionality vectors, where most of the components are 0 because only a few words do co-occur. Some techniques were developed to reduce dimensionality and to densify these word vectors. Singular Value Decomposition (SVD) (Landauer & Dumais, 1997) is such a technique, and it led to the development of Latent Semantic Analysis (Landauer et al., 1998) which is still widely used today when it comes to representing word-vectors in similarity space. The term *word embeddings* originates from the fact that words are mapped or 'embedded' into a low-dimensional similarity space (Levy et al., 2015; Levy & Goldberg, 2014).

Machine learning has recently taken computational linguistics by storm, and many algorithms were developed to automatically derive word embeddings from a given corpus. These newer algorithms are often based on neural networks, which are computing systems that were first inspired by how neurons interacted in the human brain (Rosenblatt, 1958). The great advantage of using such neural networks is that they are very flexible to the task they are trained to perform.

A neural network consists of many layers of neurons linked together, and every time a specific neuron is activated, it sends a signal to connected neurons. This signal's strength is proportional to the weight of the connection between these neurons (Goodfellow et al., 2016). There exist different ways of training a neural network to perform a given task, but the basic idea remains the same. Namely, that we provide inputs and outputs so that the network learns to match them by adjusting the weights of the many connections between the different layers of neurons as represented in Figure 4.3 (Marcus, 2018, taken from p.4). In other words,

training a neuron network is like solving an optimization problem (Smith, 2019).[66]

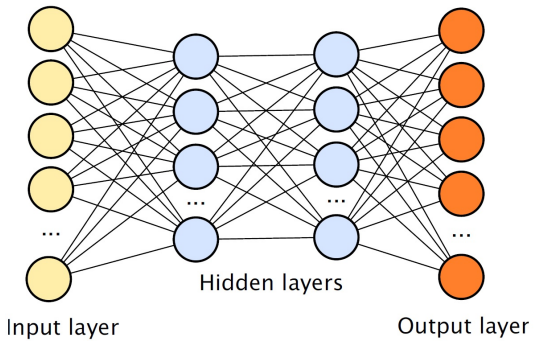


Figure 4.3 Basic representation of a simple neural network with four layers

A downside of these approaches is that in order to train these neural networks, we need a massive amount of data. On the other hand, neural word embeddings are shown to be very effective to model various tasks in computational linguistics such as paraphrase detection (Socher et al., 2011), sentiment analysis (Bojanowski et al., 2016), syntactic parsing (Socher et al., 2013), short answer tasks (Koleva et al., 2014), or sentence entailment (Sadrzadeh et al., 2018).[67]

Another limitation of the neural networks is the lack of explainability because when a neural network is trained, it does not inform us of the role of a specific node within the network, i.e., how the model's performance is linked with specific regions of the network. The opaqueness of neural networks is often problematic for those of are interested in understanding the internal mechanisms that lead to the realization of a given task (Eppley, 2014), but that does not mean that we could not still benefit from these computational models. Approaches based on

[66]The term *Deep Learning* refers to the process of training neural networks that have multiple hidden layers.

[67]I do not discuss these applications further as they are outside the scope of this thesis.

deep learning help provide an infrastructure that contributes to the understanding of linguistic processing, and the role of the linguist is to contribute to the improvement of realistic language models (Linzen, 2019).

4.1.1.1 Word2vec

In this section, I briefly present two ways to derive word embeddings.[68] One of the best-known models for deriving word embeddings is Google's word2vec Mikolov et al. (2013a,b,c). Word2vec is a three-layer neural network that is trained one word at a time to learn word representations to predict the words surrounding the input word (Lupyan & Lewis, 2017).

Two main learning algorithms are available in word2vec, i.e., two model architectures to derive word embeddings: continuous bag-of-words (CBOW) and continuous skip-gram. In the CBOW model, the target word is predicted according to its surrounding linguistic context, whereas in the skip-gram model, surrounding words are predicted from the target word (Mikolov et al., 2017). It is generally accepted that the CBOW model is faster to train, but the skip-gram model remains better with less frequent words within a particular corpus (Mikolov et al., 2013a).

4.1.1.1.1 Continuous Bag-of-words

In word2vec, the CBOW model learns word embeddings by predicting a word$_j$

[68] In this thesis, I am primarily interested in the differences in representations and training, and I will not delve into technical details nor discuss comparative results for specific NLP tasks.

from its linguistic context, consisting of a symmetrical window of words surrounding this target word. In Figure 4.4 (Mikolov et al., 2013a, Figure 1), we see that the word w_j is predicted from the previous words words w_{j-2}, w_{j-1}, and the next words w_{j+1}, w_{j+2}. These contextual windows can be large (8 words) or small (only the preceding and succeeding words), depending on the training algorithm.

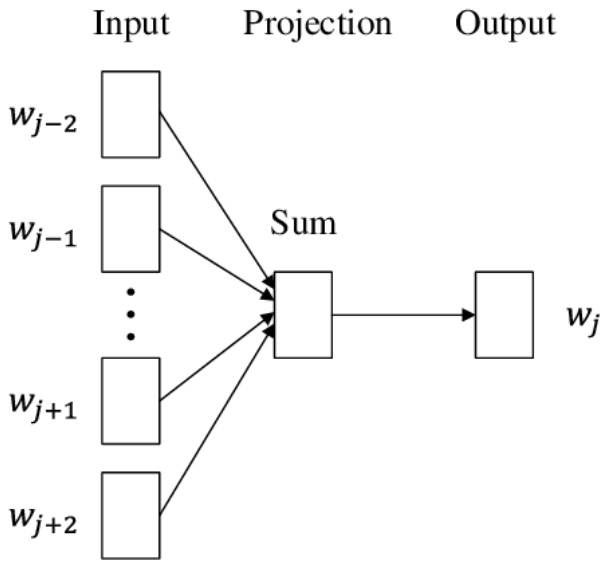


Figure 4.4 The architecture of the CBOW model

In the CBOW model, word order is not taken into account by the training algorithm, and sentences are treated as lists of words. For example, the sentence (74) would be decomposed into a list of words, and the two lists (74-a) and (74-b) would be treated the same way by the model even though the words are ordered in different ways.

(74) Kim is playing badminton.

a. “Kim”, “is”, “playing”, “badminton”

b. “badminton”, “Kim”, “playing”, “is”

During the training phase, the algorithm tries to maximize the probability of the target word conditional's log-likelihood to a given linguistic context. For example, in Figure 4.4, we use four different words surrounding the target word w_j, i.e., two words before and two words after, to derive the probability of having this target word w_j as an output. After every prediction cycle, when a prediction is not good enough, an error signal is sent to adjust the vector's values for the target word, and the components of the word-vectors are adjusted to reflect the probability of the prediction better.

4.1.1.1.2 Continuous Skip-gram

The other alternative is based on the n-gram model, which can store information about word order within a sentence (Bojanowski et al., 2016). Contrary to the CBOW model, the skip-gram model's training objective is to predict nearby words from a word (Mikolov et al., 2017) as in Figure 4.5 (Mikolov et al., 2013b, Figure 1). The continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships (Lupyan & Lewis, 2017). The model architecture is different from the CBOW because the input is a single word, but the output is multiple pairs of words.

In the skip-gram model, a sentence is decomposed into n-grams which consists of a consecutive subsequence of length n of some tokens k. For example, in (75), if we set the size of the training context to 2 and we choose the input word to be *two*, then we can form 4 different 2-grams (or bigrams) as seen in (75-a).

(75) Kimiko bought *two* big grapefruits.

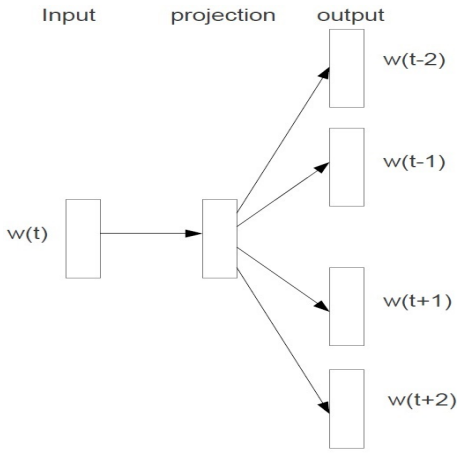


Figure 4.5 The architecture of the Skip-gram model

a. 2-grams: (two, Kimiko), (two, bought), (two, big), (two, grapefruits)

During the training phase, the word representation of the input word is used to maximize the conditional log-likelihood of the output n-grams. The input word-vector is then adjusted accordingly. At every training cycle, a different input word is used. Its word-vector is also updated from the predicted probability of the outputs until the adjustment process reaches a stable state (Gastaldi, 2020).[69]

Other models of word embeddings like GloVe have proven to perform well when measuring similarity and representing analogies (Pennington et al., 2014). However, GloVe only performs better than word2vec when the training corpora are very large; with smaller corpora, GloVe is slightly worse than word2vec (Lupyan & Lewis, 2017). Additionally, word2vec is faster to train than GloVe (Levy et al., 2015). One limitation of count-based statistics like GloVe is their permeability to the addition of new material (Lenci, 2018). When new distributional data

[69]Improved versions of word2vec now exist where the continuous skip-gram model is extended, i.e., FastText (Bojanowski et al., 2016), or even versions where training mechanism can learn richer word representations (Mikolov et al., 2017) but I will not discuss these here.

is added, we must re-train the model to integrate these new data because their word embeddings are derived from global statistics. For all these reasons, I use word2vec word embeddings for the calculation in this thesis.[70]

4.1.2 Comparing Word Embeddings: Similarity

Landauer & Dumais (1997) defined the word similarity as the distance between the vectors representing those words. Cosine similarity is the most used measure of similarity in the literature (Baroni et al., 2014b), and it is the cosine similarity used by word2vec and GloVe when dealing with analogy relations (Mikolov et al., 2013c; Pennington et al., 2014). The cosine measure is the value of the angle between the two vectors.

(76) $$Sim(\overrightarrow{w}_1, \overrightarrow{w}_2) = Cosine(\overrightarrow{w}_1, \overrightarrow{w}_2) = \frac{\overrightarrow{w}_1 . \overrightarrow{w}_2}{\|\overrightarrow{w}_1\| \|\overrightarrow{w}_2\|}$$

In (76) $\|\overrightarrow{w}_1\| = \sqrt{\sum_i w_i^2}$ is the norm of the vector. The more similar two vectors are, the smaller the angle is. The cosine measure gives out a real number from -1 to 1, 1 being perfect similarity.[71] As we see in Figure 4.6 in this hypothetical two-

[70] Another very popular word embeddings model is BERT (Devlin et al., 2018). BERT is derived differently than GloVe or word2vec, it has context-dependant word embeddings and it generally needs to be fine tuned to a particular corpus of data (see section 5.3.2.3 in Chapter 5). In this thesis, I am using a context independent approach like word2vec to illustrate how a vanilla model could be implemented without any need for fine-tuning like when using BERT. I will leave this for future work, but it would be relevant and interesting to perform the same calculations using BERT and compare the results.

[71] The possible range for the cosine measure depends on the normalisation used for the word-vectors. If the word-vectors only have positive components, then the angle between two vectors will not be greater than 90° which means the cosine measure will be between 0 and 1.

dimensional representation, the similarity between $\overrightarrow{\text{coffee}} = (2, 3)$ and $\overrightarrow{\text{tea}} = (4, 1)$ is measured by the cosine of the angle θ between the two vectors. In this case $Sim(\overrightarrow{\text{coffee}}, \overrightarrow{\text{tea}}) = 0.740$.

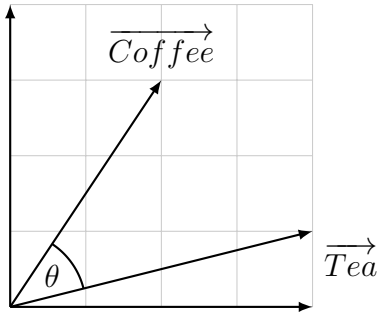


Figure 4.6 Hypothetical Vector representations for *Coffee* and *Tea*

An alternative to the cosine measure is the Euclidean distance between two vectors, i.e., the segment's length connecting their endpoints. One problem with this distance is its sensitivity to the magnitude of the vector (Baroni et al., 2014a). Consider, for example, a situation where we want to compare two-word vectors, one of which is often present in the corpus and the other that is rarely used. This computation would result in one vector being large (the one with the word with the higher occurrence) and one vector being small (the one with the rarely used word).

If we measured the Euclidean distance between the two vectors, the result would be big because of the significant difference in length between the two-word vectors. This could be problematic because we would get a low similarity because of the relative difference in the words' occurrences even though these two words might have a similar meaning. In comparison, the cosine measure is about the angle between the vectors, and their relative length does not influence it. In Table 4.1 we show different cosine similarity results computed using the CBOW and the Skip-gram

training model of word2vec with word embeddings trained on the novel Peter Pan.

Table 4.1 Cosine similarities computed using the CBOW and the Skip-gram training model of word2vec

Word 1	Word 2	Model	Cosine similarity
peter	wendy	CBOW	0.99993336
peter	wendy	Skip-gram	0.9958692
peter	temperature	CBOW	0.73825425
peter	temperature	Skip-gram	0.8839463

In this thesis, we use the cosine measure. Still, there are other alternatives like Lin (1998) who proposed a similarity metric based on information theory, and Curran (2003) who introduces other methods like the Sorensen-Dice coefficient and the Jaccard measure that originate from ecology.[72] There is also the Kullback-Leibler divergence (or KL-divergence) which is an entropy-based measure about the distance of the entropy of two probability distributions (Balkir et al., 2015). The Kullback-Leibler is used in other fields such as Cognitive Neurobiology and Predictive Processing (Friston et al., 2013, 2015) or Game-theoretical models (Skyrms, 2010; Franke & Degen, 2015).[73]

We can represent the similarities between all the words present in a corpus by using a similarity matrix. In a similarity matrix, each element corresponds to the similarity value (sim) between the word$_i$ and the word$_j$ as illustrated in Figure 4.7, where each row j and each column i correspond to a specific word embedding within the corpus. For cases where $i = j$, the similarity sim(word$_j$,word$_i$)=1. The

[72]See Kiela & Clark (2014) for a systematic study of many different similarity measures for semantic spaces.

[73]Other novel measures based on the Kullback-Leibler divergence such as Lookahead information Gain (LIG) (Aurnhammer & Frank, 2019a,b) have been proposed, but I am not going to discuss them here.

dimension of this similarity matrix is $N \times N$, where N is the number of word embeddings.

	...	word_{i-1}	word_i	word_{i+1}
...	...	...	...	...
word_j	...	$\text{sim}(\text{word}_{i-1},\text{word}_j)$	$\text{sim}(\text{word}_i,\text{word}_j)$	$\text{sim}(\text{word}_{i+1},\text{word}_j)$
...	...	...	...	...

Figure 4.7 Similarity Matrix for every combination of word embeddings

Distributional models were first developed to deal with words in isolation, but they rapidly expanded to represent not only words but whole sentences (Clark, 2015), and the cosine measure can also be used to compute the similarity between sentences. The only difference here is that the vectors are sentence-vectors rather than word-vectors. Sentence similarity is used in several applications like paraphrase detection and short answer tasks (Koleva et al., 2014), automatic summarisation (Erkan & Radev, 2004), image and web page retrieval (Li et al., 2006) and machine translation (Liu & Zong, 2004).

4.1.3 Using Word Embeddings

It is important to note that within the distributionalist perspective that emerged in linguistics, we can distinguish two kinds of distributionalism: a weak and a strong version (Lenci, 2008).

According to the weaker version, the distribution of a word's use allows extracting some paradigmatic features about the meaning of this word. In other words, by looking at the context in which a word is used, we can learn something about the

semantic and syntactic features this word possesses. Similarly, if two words are often present in the same linguistic context, we can readily assume that they also possess similar features and must have a similar meaning. However, this does not entail that these features are linked with how the representation of this word's meaning is stored within one's mind.

On the other hand, in the stronger view about distributionalism, the syntactic and semantic features extracted from the context of use are linked to how concepts are learned and categorized in the mind. When a word is encountered, it is learned from its distribution, and the information originating from this context helps the learner define and circumscribe its meaning. This strong distributionalism view is not that different from the multiple-trace memory model of Hintzman (1986) that argues that each experience encountered by a learner produces a trace in its memory, and it is this trace that associated with a concept or a memory cue. For example, we could learn the meaning of a word like *banana* by looking at one banana or by hearing someone describing one, but we could also learn it by noting the linguistic context in which it is used (Lupyan & Lewis, 2017). This view about cognitive distributionalism was espoused quite early during the development of distributional semantics, and both Latent Semantic Analysis (Landauer & Dumais, 1997; Landauer et al., 1998) and Hyperspace Analogue to Language (Lund & Burgess, 1996) were construed as a psychological and cognitive representation of words (Lenci, 2008).

It is often argued that a distributional model "offers both a model to represent meaning and computational methods to learn such representations from language data" (Lenci, 2018, p.152), but the nature of this linguistic data is rarely discussed. It is essential to understand what they represent and what their limits are (Lupyan & Lewis, 2017). In the remainder of this section, I discuss the nature of these word embeddings, and I draw an analogy between their associated similarity spaces and

the notion of conceptual spaces as defined by Gärdenfors (2004); Gärdenfors & Williams (2001).

4.1.3.1 Nature of Word Embeddings

Distributional semantics is generally based on the statistics of the context of use for a word (Turney & Pantel, 2010). If we happen to know these statistics, then we can know which word is most likely to be used given a particular context. For example, let us consider the linguistic context in (77-a), and the possible target words in (77-b).

(77) a. The men saw the chair on the ... that was glowing through the mist.
b. grass (75%), bridge (25%)

If someone had encountered this linguistic context many times during her lifetime, and if for about 3 out of 4 of these encounters, this linguistic context was completed with the word grass, then it would only be natural for this person to use *grass* to complete this sentence. This example shows that we need the co-occurrence statistics to predict the target word but that we do not need to understand why this co-occurrence existed in the first place.

Using a co-occurrence statistic to complete a sentence is very similar to what the participants have to do when performing a cloze task. In those cases, the only difference is that the participant does not explicitly have access to these statistics. They thus have to find ways to come up with their statistics of use. This is precisely what the predictability value is: the statistics of co-occurrence given

by participants that were asked to match linguistic context with target words. The question now becomes: how can these participants derive these predictability values if they do not have access to these co-occurrence statistics.

Word embeddings models are designed to feed on the natural regularities of language to capture linguistic features that represent the complex structural system of opposition that holds language together (Gastaldi, 2020). To illustrate what these structures of opposition are, we can go back to the previous example.

(78) a. The men saw the chair on the ... that was glowing through the mist.
b. grass (75%), bridge (25%)

In (78), we have only two possible continuations. This means that if *grass* is used, then *bridge* could not be used. The fact that only *grass* or *bridge* could be used is not directly related to their similarity of meaning, but it is related to the regularity of their use. These two words are thus opposed when it comes to their use within this linguistic context. This conception of word uses in terms of opposition is in line with the view that communication is a language game in the sense that meaning emerges when a game is being played between two interlocutors (Wittgenstein, 1953). In a particular language game, a speaker's move consists of selecting a word and rejecting other words, and this move puts the rejected words in opposition to the chosen one.

This relationship between the linguistic context and a continuation word is thus not really about co-occurrence, but about bi-duality, i.e. "about the relation of duality a term maintains with the dual contexts of another term" (Gastaldi, 2020, p.35). Two words are similar if they both have a similar relationship with different contexts. This relationship must be bi-dual because the relationship goes both

ways: two words are similar if they both are alternatives in a similar context, and if they both are not alternatives in other contexts. For example, *cat* and *dog* would be similar because they could be both used in similar contexts, and they also would not both be used in other similar contexts.

(79) a. I adopted a (cat/dog).

b. I went to the library to borrow a (~~cat~~/~~dog~~).

Word embeddings are an indirect way to measure this coordination between linguistic contexts and words, and it gives us an image of language use that is both abstract and autonomous from any pre-conceived analysis about the meaning of words (Gastaldi, 2020).[74] Word embeddings models that use big corpora extract these bi-duality relations and can then represent them in terms of similarities. It is important to note that these similarities are not primarily about meaning but use.

4.1.3.2 Cloze Task and Similarity Measures

As discussed in Chapter 2, similarity measures have often been linked with processing difficulty. For example, the results of Roland et al. (2012) tend to show that more similar words were processed more rapidly, even in a context where the more similar words were supposedly less predictable. From their results, they argued for similarity over purely predictability-based approaches.

---

[74] In his paper, Gastaldi (2020) describes the bi-duality relations in terms of Hjelmslev's and Saussure's conception of language, but it would be beyond the scope of this thesis to discuss this here.

Measuring the effects of similarity when processing a word is not the same as using similarity to predict an upcoming word directly. In a processing task, participants are given a linguistic context and the target words, and what is measured is the relative processing time between different alternatives for this target word. This way, similarity measures between the prior context and the target words can be compared for the different alternatives.

In a predictive task, the goal is to predict an upcoming word, and this involves a different use for similarity measures. To illustrate how similarity measures could be used to derive a prediction, we can build a toy model in which the similarity measure between the target word and the rest of the sentence would drive the predictive process. In (27-a), reprinted here as (80), we can see that *spear*, *sword*, *machete* and *rock* are all probable continuations for this truncated sentence. The upcoming word that best fits the context will be the most similar to the context. In this toy model, we are not considering meaning compositions, which means that the truncated sentence is processed like a list of words. In this case, we have the words *soldier*, *jabbed*, *angry* and *lion* that needs to be accounted for.[75]

(80) The soldier jabbed the angry lion with a...

To measure its similarity with the different possible alternatives, we first take the average of the word-vectors present in this context (Mikolov et al., 2013a) and then compute the words that are the most similar with this average context. Table 4.2 presents a list of predicted words and their respective similarity value as computed using the word2vec-google-news-300 and the GloVe-twitter-200 pre-trained word

[75]Very common words like the short function words *the* and *a* are referred to as *stop words* and they are generally filtered out when processing natural language data. Depending on the context, stop words could play an important role when deriving a linguistic prediction, but taking them into account is a challenge laying outside the scope of this thesis.

embeddings.[76]

Table 4.2 Words that are most similar to the truncated sentence in (80)

Training Corpus	Words	Similarity Measures
GloVe-twitter-200	*enraged*	0.520
	wounded	0.492
	man	0.488
	shouted	0.485
	yelled	0.485
word2vec-google-news-300	*enraged*	0.546
	solider	0.509
	irate	0.507
	policeman	0.500
	jabbing	0.492

In Table 4.2, we see that the predicted words make little sense. These results show that we cannot solely rely on similarity measures to find the best possible continuation for a truncated sentence. In turn, we can conclude that bi-duality relations are not enough by themselves to model linguistic prediction.

This poor result should not be too surprising because trying to predict the next word using an average context would be like trying to find a word that has a similar bi-dual relationship with the context expressed in the truncated sentence.

[76]By taking the average of the word-vectors, we are not taking into account the syntax of the sentence. This lack of consideration for the syntax traces back to the first implementation of the Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997), where sentences were represented as the sum of the word-vectors composing them. Under the LSA perspective, it was thought that the syntax had only a negligible contribution to the information retrieved from the sentence (Gastaldi, 2020). This 'omission' is undoubtedly one of the most prominent critics against such models of distributional representations (Turney & Pantel, 2010).

This is very different from when participants are asked to perform a cloze task.

Consequently, we must consider other relevant linguistic information if we want to use word embeddings to compute a linguistic prediction. For example, we have to consider that we are looking for a specific word that will fit well with the whole context and fit well with the different granularities of linguistic information.

For example, the fact that the predicted word follows the preposition *with* is a strong indicator that the target word will be an instrument of some sort.[77] For the predictive model to consider this, we have to account for the different levels of meaning within the sentence. The solution to this problem is to compute similarity measures between meaning compositions rather than between words. This way, we could account for the different levels of meaning within the sentence and the fact that it is hierarchically parsed (Ding et al., 2017).

4.1.3.3 Cloze Task and Semantic Networks

The fact that we can generally predict the next word of (81) to be *bread* is linguistically motivated, i.e., it is not that *bread* is similar to the rest of the words in the sentence, it is because the meaning expressed by the word *bread* fits well with the meaning composition expressed the first part of the sentence. However, before discussing the nature of these meaning compositions, we must find a way to transpose similarity measures into measures that could fit with the cognitive process associated with cloze tasks.

---

[77] Instrument in the sense of something that a soldier could jab a lion with.

(81) I went to the bakery for a loaf of ...

When a participant reads or processes a sentence during a cloze task, this person is also incrementally building a mental representation of this sentence's meaning. Although essential to understand the cognitive processes that are taking place in a communicator's mind, I postpone the discussion on incrementality to Chapter 6 to focus more on the nature of this mental representation in Chapter 4 and Chapter 5. In these chapters, I am primarily interested in discussing the principles behind the derivation of a linguistic prediction, and these principles are independent of the incremental nature of linguistic processing.

To predict the upcoming word, the participant must weigh all possible continuations and choose the one word that best fits this mental representation. One way to illustrate how these possible continuations are obtained is to represent them in an activation-based semantic network node. A semantic network is a network where words are represented as lexical units related to other words via semantic relations as illustrated in Figure 4.8. One of the most famous databases of such semantic relations is WordNet (Miller, 1995).

In Figure 4.8, we see that *mammal* is related to *animal*, and that *animal* is in turn related to *fish*. This particular network does not represent the weight of the connection between the lexical units. However, it still allows us to see that *mammal* sits closer to *animal* than to *fish*. In terms of activation levels, this means that if someone utters *mammal*, the co-activation of *animal* will probably be stronger than the one for *fish*. Activation-based semantic networks allow us to represent the strength by which two lexical units are connected. Most importantly, it allows us to derive the activation level at a given node in terms of the activation of other nodes. In a semantic network, not every word is related to all the other

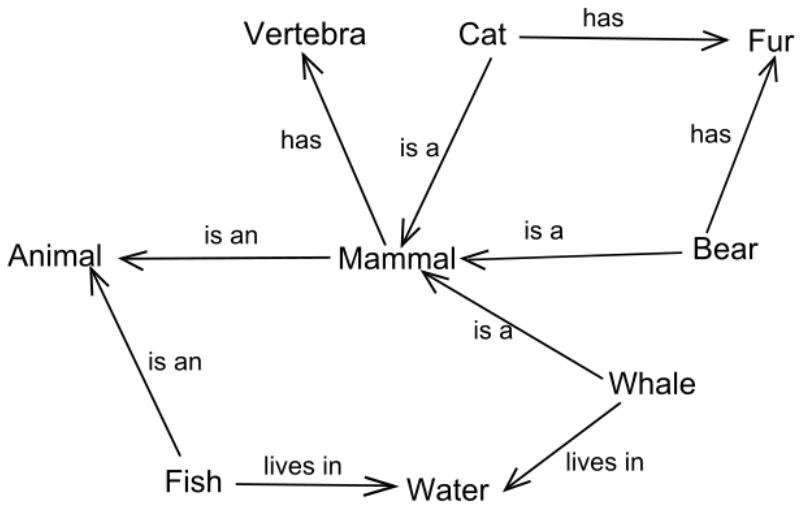


Figure 4.8 Example of a semantic network centered around the word *mammal*

words, and it is possible to have only a few selected words activated while most of them are not activated.

Activated-based semantic networks could help represent the constraining process happening during the derivation of the prediction. To achieve this, we can use an association game where someone is asked to match words together like in (82). Here, it is clear that the word *elephant* would not be associated with the word *car*.

(82) *car*: *wheel*, *motor*, ~~*elephant*~~

It is possible to transpose this idea of associations between words to the cloze task where a given mental representation corresponding to the meaning of a sentence could only give rise to the activation of certain specific target words. In a cloze task, the best word for the continuation of a sentence is selected only from the words that are minimally activated, not from all the words that the participant

knows. This kind of activation-based semantic network allows us to model linguistic prediction as a spreading-activation process (Collins & Loftus, 1975). However, since our primary material is coming from word embedding models, we must find a way to transpose them to use similarity measures in this activation-based semantic network.

4.1.3.3.1 Similarity Spaces are Conceptual Spaces

In order to transpose word embeddings into activation-based semantic networks, I am using the notion of *Conceptual Space* (Gärdenfors, 1996, 2000; Gärdenfors & Williams, 2001; Gärdenfors, 2004, 2014). A conceptual space is a geometric framework "designed for modeling and managing concepts" (Gärdenfors & Williams, 2001, p.1). One advantage of conceptual spaces is that their geometric nature allows them to flexibly interact with different kinds of representations, enabling them to act as a sort of Lingua Franca of knowledge representation (Lieto et al., 2016), and this is precisely what we need.

A conceptual space is a multidimensional feature space where a concept's position is based on various quality dimensions (Khater & Tawfik, 2009). These quality dimensions correspond to the different ways in which concepts are judged to be similar (Gärdenfors, 2004). Analogously to what I described before, conceptual spaces can also be viewed as an n-dimension vector space where concepts are represented as regions of this space. Conceptual spaces are represented with a Voronoi diagram, a plane divided into different contiguous regions corresponding to these quality dimensions. As an illustration, we see that in Figure 4.9 (Petitot, 1988, taken from p.69), the stop consonants are spatially positioned using two axes: voicing and place of articulation.

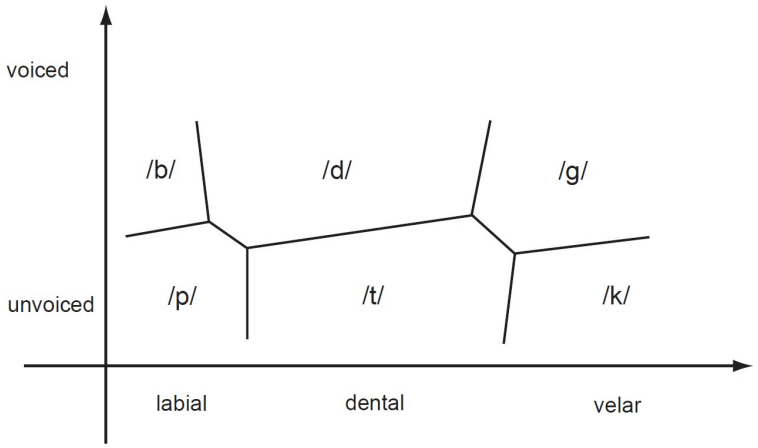


Figure 4.9 A Voronoi diagram of the boundaries of stop consonants

In Figure 4.9, the regions surrounding each consonant represent the possible variability in the pronunciation of such consonants, i.e., the prototypical pronunciation would be centrally located within a region (Gärdenfors & Williams, 2001). Additionally, the boundaries represent the transition points between those regions, i.e., when the place of articulation changes from labial to dental, then the /b/ becomes a /d/.

Positioning a concept in a conceptual space is equivalent to categorizing it in terms of its qualities. One way to do this is to measure its similarity with the prototype of this category. For example, a concept X belongs to the same category as the prototype Y as long as there is not another prototype that is more similar to it (Gärdenfors, 2004). Similarity relations are crucial for conceptual spaces because they represent an essential aspect of the differences between concepts (Gärdenfors & Williams, 2001). This way, differences between concepts are calculated as the distance between two concepts on the Voronoi plane (Gärdenfors, 1996). As in any geometrical plane, measuring a distance could be done using different methods (e.g., Euclidean distance, cosine distance).

One limitation of this approach is that we usually need to involve psychological experiments about similarity judgments to construct these conceptual spaces. However, with the advent of modern computational approaches, similarity spaces could be directly derived from word embeddings and machine learning methods (Douven & Gärdenfors, 2020). Once we have derived these similarity spaces, we can use transformational matrices to map one space to another (Lieto et al., 2016).

One crucial difference between similarity spaces for word embeddings and conceptual space is the fact that the latter has n-dimensions that correspond to quality dimensions (Gärdenfors, 2004). On the other hand, dimensions of similarity spaces correspond to unidentified bi-dual regularities between a target word and the context it is used in, as we described in Section 4.1.3.1. Even if word embeddings dimensions have been described as not being explicitly meaningful (Lieto et al., 2016), they do capture a regularity of structures that may encompass a regularity of meanings. The model presented in this chapter is compatible with the idea put forward by Gärdenfors (2004) that judgments of similarity are related to cognitive structures of knowledge.

4.1.3.3.2 From Word Embeddings to Semantic Networks

When transposing the similarity spaces into semantic networks, we can use the size of a region as the threshold of similarity between two words, i.e., if a word w_1 is equally similar to two different prototypes $p2$ and $p4$, then it will sit precisely at the boundaries between these two regions as in Figure 4.10 (Gärdenfors, 2004, adapted from their Figure 6).

If a particular region is large, then its prototype will have a more significant in-

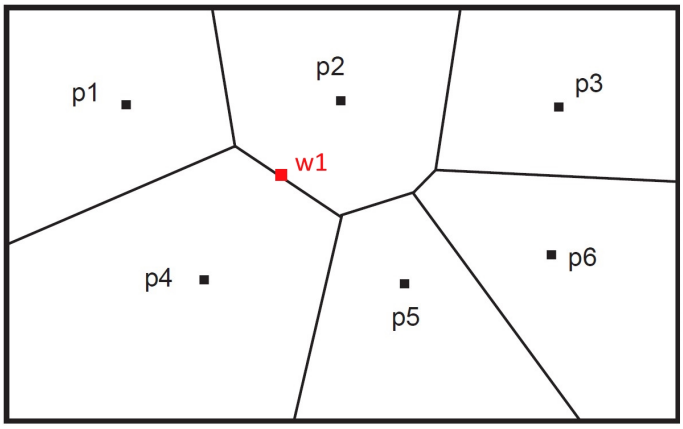


Figure 4.10 Radius around a prototype in a Voronoi diagram

fluence on its neighborhood than a prototype from a smaller region (Gärdenfors & Williams, 2001). This geometrical notion of similarity could be helpful when dealing with non-monotonic effects related to context updating. For now, geometric considerations are not essential because we are primarily interested in the activation-based semantic network, which, by themselves, are not geometric spaces anymore.[78]

To transpose word embeddings into an activated-based semantic network, we use the fact that similarity measures correspond to the weight of the connection between words. This transposition in terms of spreading activation improves lexical processing models based solely on similarity measures (Rotaru et al., 2018).

For example, suppose the similarity measure between $Word_a$ and $Word_b$ is 0.9, and the similarity measure between $Word_a$ and $Word_c$ is 0.5. In that case, the weight of the connection between $Word_a$ and $Word_b$ will be greater than the weight between $Word_a$ and $Word_c$. This transposition of similarity measure into

[78]Geometric considerations of the similarity space could play a role when integrating contextual influences, and I discuss this issue in Chapter 6.

connection weight is illustrated in Figure 4.11 where the thickness of the connection represents the weight of the connection, and the similarity measure is taken to be the Euclidean distance.

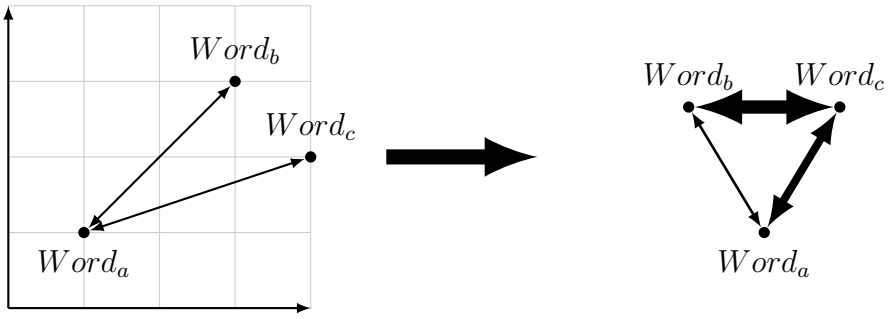


Figure 4.11 Example of a transposition of similarity spaces into connection weight in 2D

Once the weight of the connection has been derived from the similarity measures, we can use this weight to measure the activation level of $Word_b$ given $Word_a$. Proceeding this way, we are assuming that the connection weight between two words is bi-directional. This feature comes from the fact that similarity measures are symmetrical. Even if this bi-directionality is not a problem at this point, it might play a role in some contexts. For example, if we are asked to name the first words that come to mind when we hear (83-a). Then many participants could answer *de Janeiro*. Under this activation-based semantic network account, this answer would mean that the strongest connection weight from *Rio* was with *de Janeiro*, which in turn would imply that their similarity was the highest. However, it is also possible that when the first word is (83-b), then the answer is *Rio*, which would again mean that the highest similarity was between *Rio* and *Carnival*.

(83) a. Rio

b. Carnival

Those two results would thus contradict each other, and it suggests that bi-directional or symmetrical relations like similarity might not be the best to use when it comes to transforming them into semantic networks. However, in this thesis, I use similarity relations between word embeddings because this unidirectional aspect of similarity becomes relevant only when considering incremental models, which is not the case here.[79]

After transposing all the word embeddings into a level of activation, we end up with a graph-like structure linking all the words to their nearest neighbors like in Figure 4.12.

When the interpreter processes a specific word, it is as if we are sending a signal through a graph like in Figure 4.12. The signal is maximal at the beginning, and it is then divided into different branches. Every time the signal encounters a new branch in the road, the signals become weaker, which is also indicative of the relative activation between words, i.e., if the connection weight between $Word_a$ and $Word_b$ is stronger than the one between $Word_a$ and $Word_c$, then $Word_b$ will be more activated than $Word_c$ which implies, in turn, that the other words that are related with $Word_b$ will also be more activated than the one linked with $Word_c$. This hierarchy of activation goes on up until the signal dissipates and becomes negligible.

From there, we can compute the activation level of a target word with respect to other words' activation. Going back to our previous example *The soldier jabbed the angry lion with a ...*, if we consider that all the words that are expressed in

[79] Another solution to this caveat would be to use an unidirectional relationship measures between word embeddings like the KL-divergence I mentioned earlier.

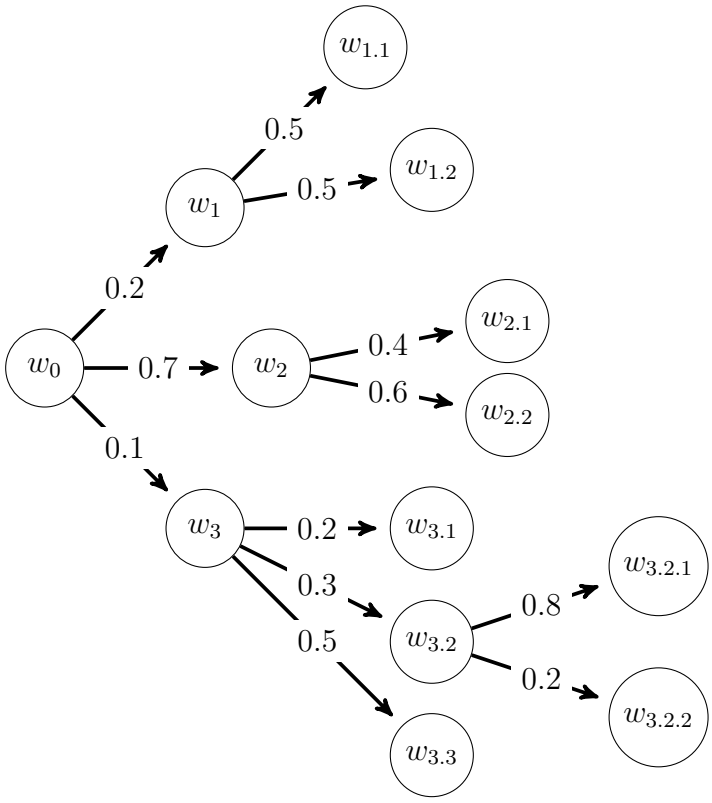


Figure 4.12 Example of an activation-based graph with normalized connection weight

the truncated sentence are equally fully activated (e.g., 1.00), then we can send this signal through the graph to get the most activated words.

In this section, I introduced word embeddings, showed different approaches to derive them, and explained that comparing word embeddings could lead to the computation of useful linguistic information. I then discussed the possible parallel between the similarity spaces described in distributional approaches and the conceptual space described in cognitive linguistics. Finally, I argued that similarity measures could predict the next word, but that we first had to transpose similarity measures into an activation-based semantic network. In the next section, I present the relevant representational levels of the meaning compositions and explain how

to compute similarity measures for every one of them.

4.2 Compositional Units

The need to have different kinds of linguistic units has already been assessed empirically in Chapter 2. We also discussed it in Chapter 3 when I argued that language use was about meaning composition and combining linguistic units into larger units of meaning. In this thesis, I consider four different representational levels of meaning related to the derivation of a linguistic prediction. After assessing their respective nature and computed their contribution separately, I discuss how to combine them to derive the linguistic prediction.

I have already mentioned that word embeddings are the primary data used by this model of linguistic prediction. However, one major problem of word embeddings is that they cannot naturally deal with the hierarchical features involved with meaning compositions, and this has to do with the way they are learned (Marcus, 2018).[80] The features used during the training process of the word embeddings are non-hierarchical, which means that any information about the way the meaning of two words is composed together is not taken into account during the training phase. One way to resolve this problem would be to modify the training algorithm so that the word embeddings could intrinsically consider meaningful hierarchical features.

---

[80]The importance for distributional representations to take into account compositions of words instead of a series of single words has already been the focus of a debate between supporters of connectionism and supporters of classical symbolic architecture (McClelland et al., 1989, 1986; Fodor & Pylyshyn, 1988; Chalmers, 1993). See Doumas & Hummel (2012) for more information about the differences between these two kinds of architectures.

In this thesis, I am instead proposing to treat meaning composition in terms of different combinations of word embeddings, each corresponding to a different representational level of the composition. Much like the lexical meanings stored in memory that we combine to construct the meaning of a multi-word utterance (Hagoort, 2020), word embeddings can be used as building blocks to derive a more complex composition of meanings. Using a compositional function to combine word embeddings is in line with the many hybrid symbolic-connectionist models developed in recent years (Doumas & Hummel, 2012; Baroni, 2013; Clark, 2016).

In the remainder of this section, I describe four representational levels associated with meaning compositions: I explain their respective contribution to linguistic prediction, and I argue that each representational level lives on parallel similarity spaces, which have all to be considered when deriving a linguistic prediction.

4.2.1 Composition of Word-vectors

The three requirements that a compositional model must meet are the following Martin & Doumas (2020, p.5): 1) we must have representations that give information about the state of the world, 2) we must have a compositional mechanism by which new structures are inferred from existing structures, 3) we must have a compositional mechanism that is independent of the derivation of the representational elements.

The first requirement is readily met by using word embeddings because these representations of lexical meaning are used to describe a state of the world. As for the second requirement, the idea is that once we represent the meaning of a word as a vector, we then have to combine different word-vectors to build a

representation for the meaning of a whole sentence (Clark, 2015). The Principle of Compositionality tells us that the meaning of a sentence is a function of its parts and how they are combined (Pelletier, 1994), and it is possible to write down the representation of a sentence or a constituent as a function of different word embeddings. Various approaches to combine word-vectors into sentence-vectors have been proposed by numerous scholars (Baroni et al., 2014b; Clark et al., 2008; Clarke, 2012) and many of them are based on the multiplication of word embeddings (Mitchell & Lapata, 2010; Coecke et al., 2010).

In their article, Mitchell & Lapata (2010) compared nine different compositional functions, and they concluded that 'simple multiplication' was better suited for sentence similarity tasks. The model proposed by Coecke et al. (2010) takes into account the structure of the composed sentence. The idea behind their approach is that "syntax drives the compositional process" (Clark, 2015, p.26) and that the composition follows the rules of pregroup grammar (Lambek, 2008).[81]

The major problem of the compositional approaches based on multiplication is that the similarity will depend equally on all the words present in the composition (Doumas & Hummel, 2012). For example, if we compare *red dog* and *red house*, the similarity will be as large as the one between *dog* and *house*. Models based on multiplication fails to grasp that the compositions *red dog* and *red house* have more in common than *dog* and *house* because these former are part of the same set of things that are red. Multiplicative approaches are thus incompatible with the third requirement from Martin & Doumas (2020), which was that new representational structures are inferred from the existing structures.

This third requirement stipulates that the compositional function should not mod-

[81]For a much more detailed description of this model, see Grefenstette et al. (2014); Grefenstette (2013); Sadrzadeh et al. (2013); Sadrzadeh (2016).

ify or affect the constituents' meaning within the composition. Multiplicative composition is thus problematic because it precludes any kind of generalization. After all, the representation of a compound will mostly depend on one word and not the composition as a whole (Doumas & Hummel, 2012). In this thesis, instead of multiplying vector embeddings, I am modeling compositions by adding features present at different representational levels.

Even though some proposed a hybrid approach with both multiplicative and additive combination (Baron & Osherson, 2011), vector addition has been described as a better way to implement the derivation compositional elements because it fulfills the three requirements for compositionality (Doumas et al., 2008, 2017, 2018) and because it sits closer to the way neurophysiological computations work (Calmus et al., 2020; Martin & Doumas, 2020). Even though the specific details regarding the link between this hierarchical representation and cortical representations remain to be specified (Martin & Doumas, 2017), it does open the door for a more mechanistic modeling of semantic composition (Martin, 2019).

In this section, I focus on one particular kind of additive model motivated by the temporal synchrony of neurons firing (Hummel & Holyoak, 2003, 2005). The basic idea behind this approach is that when two words are combined, the representations of both words are added together similarly to when two neurons are co-activated and both fire at the same time (Doumas et al., 2008).

This model called Learning and Inference with Schemas and Analogy (LISA) is a hybrid symbolic-connectionist model that codes relational structure and can represent both objects and relational roles as patterns of activation over units representing semantic features (Hummel & Holyoak, 2005, p.154). This model was first developed to help solve the incapacity for traditional connectionist networks to represent relational structures explicitly (Hummel, 2011). For example, in a

traditional connectionist setting, the relational structure of *loves(Kyle,Lory)* is represented as a whole, and it does not contain the explicit contribution from either the role or the filler.

Instead of a purely connectionist model, we can use a symbolic structure in a connectionist setting to model the binding relations between a role and a filler (Hummel & Holyoak, 2003) while maintaining the independence of the components that are composed together (Martin & Doumas, 2020). This hybrid model remains unique because of its ability to "provide an overarching vista on theoretical, computational and experimental research on syntactic and semantic composition." (Martin & Baggio, 2020, p.6).

In the LISA model, a proposition like *knows(Sally,loves(John,Sally))* is hierarchically represented using four levels (Hummel & Holyoak, 2003, 2005; Doumas et al., 2008; Martin & Doumas, 2017) as we can see in Figure 4.13 (Hummel & Holyoak, 2005, from their Figure 1). The lowest level consists of different features related to the given words, i.e., Semantics Units. For example, in the left-most part of Figure 4.13 we can see that both *Sally* and *John* are connected with the semantic features 'human' and 'adult'. The second level is for the localist Predicate-Object units for individual objects (Sally and John) and predicates (lover and beloved), i.e., PO units or lexical level. The third level is for sub-propositions that bind objects and predicates (John is the lover). It is about local role-binding, i.e., RB units. Finally, at the top level, we have the complete proposition, i.e., P units. For cases where we have a complex proposition like *knows(Sally,loves(John,Sally))*, then we have to represent the P unit of the lower order proposition, i.e., *loves(John,Sally)*, and then embed it as an argument of the higher-order P unit, i.e., *knows(Sally,...)* (Doumas et al., 2008).

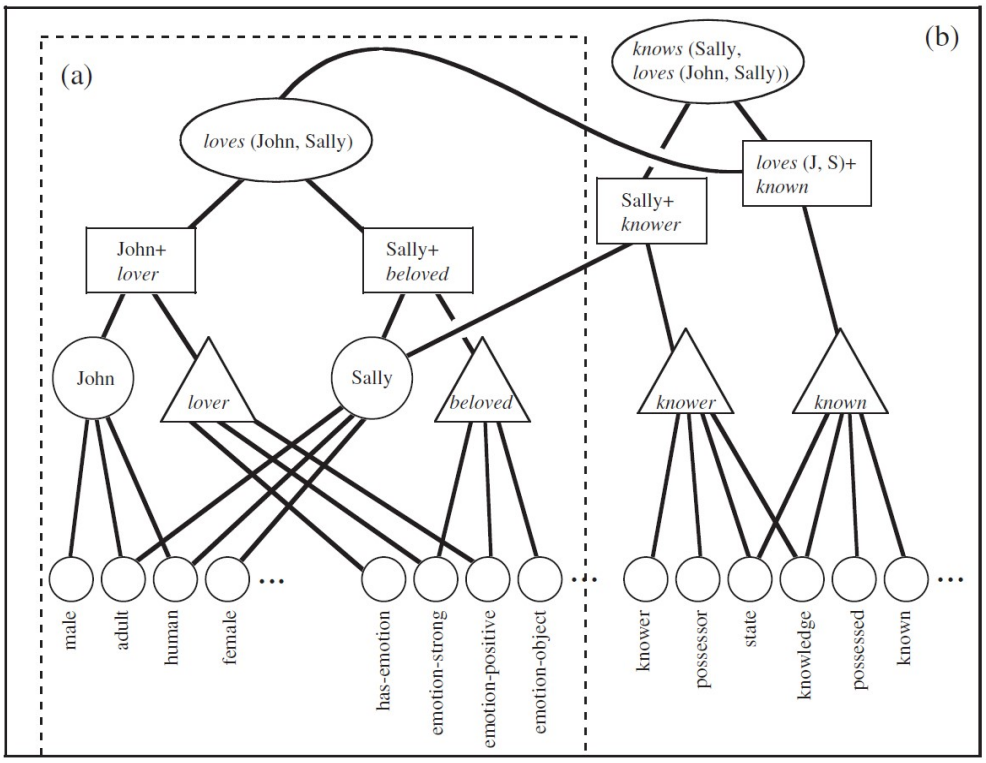


Figure 4.13 Representation of the proposition *knows(Sally,loves(John,Sally))* in the LISA model

According to this view, the meaning expressed by a proposition is thus conceived as the addition of all the units created by this proposition. In terms of neuron firing, a proposition can be represented as the sum of the activation of semantics features, lexical units, PO units, and RB units. To better visualize this, we can use a time-series illustration of the representation of a particular proposition like in Figure 4.14 (Hummel & Holyoak, 2003, from their Figure 3). It is important to note that in LISA the units corresponding to *Bill* and the one corresponding to *lover* fire synchronously, but in asynchrony with those for *Mary* and for *beloved* (Hummel & Holyoak, 2003).

The fact that in LISA, we have a systematic asynchrony, i.e., some units fire in

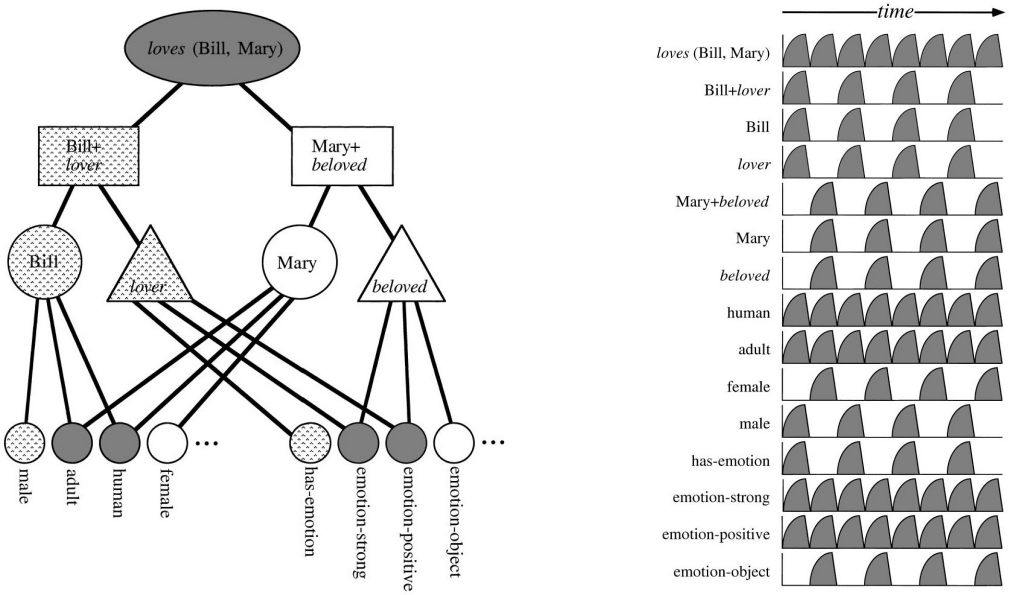


Figure 4.14 Left-side: Representation of the proposition *loves(Bill,Mary))*in the LISA model. Right-side: Time-series illustration of this representation

synchrony, some do not, is problematic in the sense that it seems to imply that the predicate units and the object units are of two different kinds (Doumas et al., 2018). This assumption is made because the only way to differentiate a predicate and an object that fires simultaneously would be to have non-overlapping semantic features for both of them (Doumas et al., 2008). To remedy this limitation, a new asynchronous model was developed to represent this temporal firing of different units: Discovery of Relations by Analogy (DORA) (Martin & Doumas, 2017). Although it is systematically asynchronous, DORA is a generalization of LISA (Doumas et al., 2008), and it is still bound to the same requirements we already discussed for a compositional model. The general hierarchical structure of a proposition is the same in LISA and DORA. Still, in the latter, the asynchronicity means the temporal firing is different, as shown in Figure 4.15 (Doumas et al., 2008, from their Figure 3).

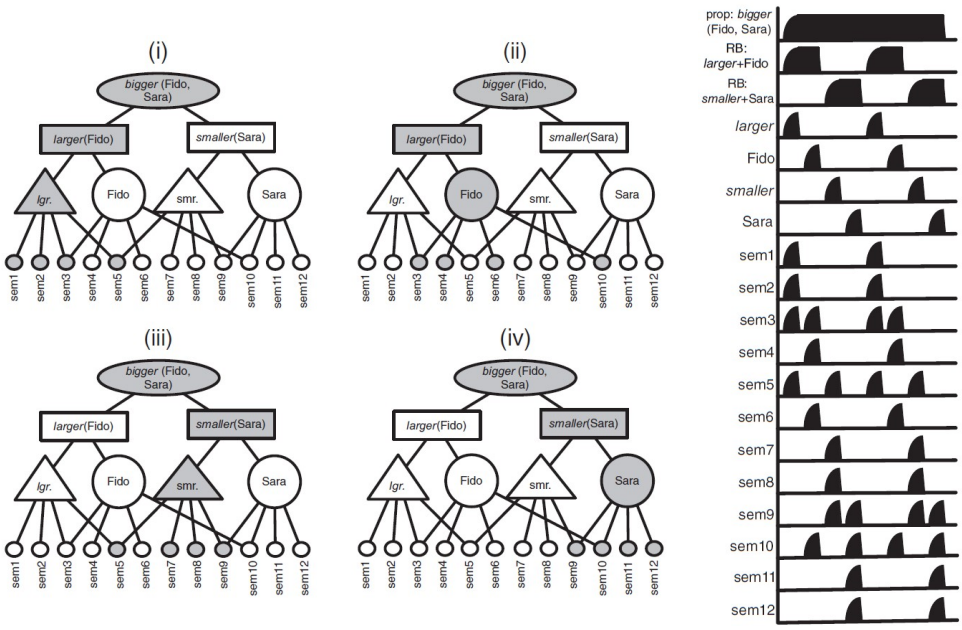


Figure 4.15 Left-side: Representation of the proposition *bigger(Fido,Sarah)* at four different time in the DORA model. Right-side: Time-series illustration of this representation

In this asynchronous binding, the PO units for *larger* fire first, followed by the PO units *Fido*, and so on up until all the PO units have fired. The distinction between synchronous and asynchronous binding is relevant primarily when we are taking into account incremental processing. Because the model presented in this chapter is not incremental, I will not further emphasize LISA and DORA's differences.

Crucially, both LISA and DORA support the idea that a multi-place proposition is represented as a series of single-place predicates that are all linked to each other, and they both posit the necessity to represent meaning composition in terms of different hierarchical levels (Martin & Doumas, 2017). The LISA and DORA framework represents compositional units as a hierarchical structure of activation comprising four levels. The meaning of these compositional units is represented as the addition of the contribution from these four levels. In the remainder of

this section, I show how to adapt this proposed representational structure to be integrated within a model of linguistic prediction.

The main idea is to transpose this representational structure into the activation-based semantic network we previously described. This way, we would consider the meaning of a composition of words by summing the contributions from the different representational levels: semantic features, lexical units, RB-units, and P-units. In other words, when predicting the next word of a sentence, we are taking into account the similarity measures between all the units that are activated at these four levels. To illustrate this process, we can transpose the representational levels as depicted in LISA for the sentence *The cup is bigger than the ball* (left-side of Figure 4.16 (Doumas et al., 2018, Figure 2)), into a representational model for a truncated sentence (right-side of Figure 4.16).

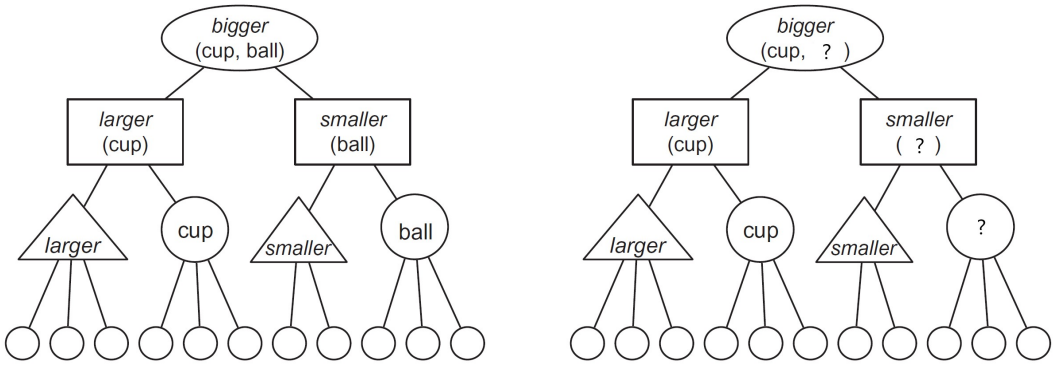


Figure 4.16 Left-side: Representation of the proposition *bigger(cup,ball)* in the LISA model. Right-side: Representation of the truncated sentence *bigger(cup,...)*

In LISA and DORA, when a proposition is activated, it produces a systematic firing pattern originating at the P-level and going down to the semantic features level (Martin & Doumas, 2019). The time-series of this cascade of firing was depicted in 4.15. When it comes to using this hierarchical approach to represent the contribution from the compositions of word embeddings, it is slightly different

because of the incrementality of linguistic processing.

As explained before, when we encounter a truncated sentence, the first level that is activated is the lexical level and not the level of the proposition. However, because we are not taking incrementality into account in the model we are presenting in this chapter, we are considering the respective contribution from these levels to be independent of time, which means that we can simply add up the contribution at every level to retrieve the linguistic prediction. In the remainder of this section, I explain the contribution from the four levels by presenting each hierarchical level separately and explaining how to measure their similarity.[82]

4.2.2 Lexical units

The predicated-object (PO) units represent the individual predicates and the objects present in the composition (Doumas et al., 2017; Martin & Doumas, 2019). In this thesis's model, I use the term lexical units instead of PO-units to underline the fact that these are directly represented as word embeddings.

To derive the lexical level contribution, we use the similarity matrix we already constructed for the lexical level, and we compute the most activated continuations. For the example at hand, we derive the most activated lexical units for the three PO-unit involved in *bigger(cup,...)*, namely, *larger*, *smaller* and *cup* as depicted in Figure 4.16. From there, we add the contributions from these three words to

[82]In this thesis, I deliberately chose to associate these four representational units with the same word embeddings. In other words, I assume these four units can be derived from the same basic conceptual space. In future work, it would be interesting to associate different word embeddings with different representational units and compare the results with the vanilla model I present in this thesis.

come up with a linguistic prediction at the lexical level. This operation is very similar to what we already described in section 4.1.2.

4.2.3 Semantic Features

Semantic features act as distinguishing features between different concepts, and they have been useful to conceptually represent concepts and categories of concepts (McRae et al., 2005). In LISA, the semantic features are treated separately from the higher level of representations which means that the semantic units are the same whether related to a predicate or an object (Hummel & Holyoak, 2003). An example of different semantic features that are attached to two words is depicted in Figure 4.17. In this particular case, we see that the semantic feature *big* is shared both by *truck* and by *elephant*. In terms of activation levels, Figure 4.17 (Doumas et al., 2008, Figure 1) tells us that if both *truck* and by *elephant* are being activated, then the semantic unit for *big* would be activated twice.

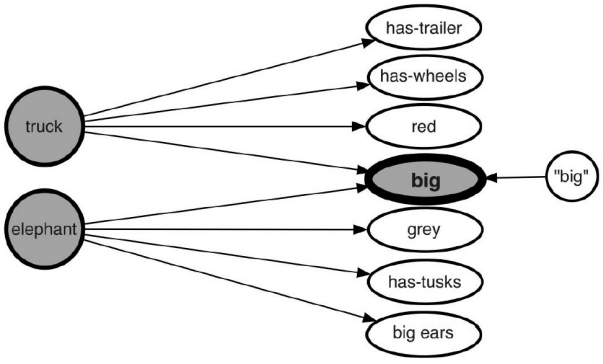


Figure 4.17 Representations of the semantic features of *elephant* and *truck*

Semantic units are activated when the word they are categorizing is activated,

but we could also determine which word is activated the most by a given set of semantic units and their respective activation level. This idea is tantamount to what we did in the previous section when we computed the best continuation as the highest activated word within the set of all the possible continuations. The main difference here lies in the fact that semantic features are not derived from computational methods based on word embeddings.

Concretely, we can show how it is done by going back to a proposition like *bigger(cup,?)*, where the last word has been removed. The first thing to do is to write down the semantic features for all lexical units present in Figure 4.16, i.e. *larger*, *cup*, *smaller*.

(84) a. *larger*: size, attribute, high
b. *cup*: dish, mug, kitchen, handle, breakable,..
c. *smaller*: size, attribute, low

In this particular example, I used *empirically derived* semantic features from McRae et al. (2005) for the word *cup*, but I had to simplify the form of the semantic features so it would be easier to compare using a similarity matrix. For example, I adapted *is_ breakable* to *breakable*. For the words *larger* and *smaller*, the semantic features are similar because both are attributes of size, one having a higher value than the other.

Once we have the semantic features associated with the words present at the lexical level, we can transpose these semantic features into word embeddings and compute the most activated word embeddings using the same procedure as the one used for the lexical level.[83] From the semantic features that are activated from

[83]At this point I do not differentiate the relative activation between features associated

the words expressed in the truncated sentence, we derive a set of pre-activated semantic features, and we can map these semantic features into lexical items using the same set of semantic features production norm from McRae et al. (2005) or use other sources for semantic features like Wordnet (Miller, 1995; Fellbaum, 1998).[84] The semantic-features units act as a constraint on the number of possible continuations. In other words, the activation of the semantic features associated with the lexical units indirectly triggers a spreading activation at the lexical level.

4.2.4 RB-units

The RB-units are representing the role-binding units or the 'sub-propositions' expressed by the full proposition. These RB-units link a single role (or predicate) with its argument to form specific role-argument (or role-filler) pairs (Doumas & Hummel, 2012; Doumas et al., 2018, 2017). For example, in the proposition *loves(John,Mary)*, there is one binding of the features of *John* to the features of the *lover*, and one binding of the features of *Mary* to the features of the *beloved* (Holyoak & Morrison, 2005). Binding is the attribution of the filler to the argument role, and it is part of the compositional process (Martin & Doumas, 2020).

To compute a prediction at this level, we have to measure the similarity between

---

with the same word and I consider that every one of them is equally activated. For more information about different measures of informativeness of semantic features, please see McRae et al. (2005).

[84] In this thesis, I used both McRae et al. (2005) and Wordnet to define the semantic features that were relevant for a given sentence. It would be beyond the scope of this thesis to discuss in which ways these two sources are obtained, but the important thing to keep in mind is that once these semantic features are transposed into a semantic network, every feature behaves the same way, i.e. every semantic feature will activate other features that are similar to it within the conceptual space.

the RB-units. This similarity is measured as the weighted average of the similarity between the words within the sub-propositions (Martin & Doumas, 2020).[85]

(85) $cos(A(w_1, w_2), A(w_3, w_4)) = (cos(w_1, w_3) + (1 - n)(cos(w_2, w_4))$

In (85), $A(w_1, w_2)$ is the composition of w_1 and w_2, and n is a weighting parameter which is set to 0.5 (Martin & Doumas, 2020). In this model, the predicates of the composition w_1 and w_3 have a greater role in the global similarity between two propositions. This way of measuring the similarity between two compositions allows us to compare the structure of the composition and not only the words constituting it as it differentiates the predicates and the filler. Consequently, (86-a) are more similar to (86-b), than to (86-c) even though the two fillers are the same because the predicates have more influence on the meaning of the composition.

(86) a. *snores(Carl)*
b. *snores(Jackie)*
c. *run(Carl)*

When it comes to linguistic prediction, we can use (85) to predict a missing word from one of the propositions. For example, if the second proposition's filler is missing, we can find the word w_4 to maximize this similarity measure. In order to do so, we have to construct a similarity matrix at the RB-level by computing the similarity for all the potential compositions using the word embeddings and the

[85]Martin & Doumas (2020) used this definition to show that human similarity ratings were corresponding better with the DORA compositional mechanism than with a compositional process involving the tensor product.

formula in (85). This similarity matrix is shown in Table 4.3. In this case, instead of the similarity value between two word embeddings w_1 and w_2, the similarity value (sim) is measured between two compositions $A(w_1, w_2)$ and $A(w_3, w_4)$. The dimension of this similarity matrix at the RB-level is $N^2 \times N^2$ where N is the number of word embeddings, and N^2 is the number of possible binary compositions between these word embeddings.

Table 4.3 Similarity Matrix for every binary combination of word embeddings

	$A(w_i, w_{j-1})$	$A(w_i, w_j)$	$A(w_i, w_{j+1})$
...	...	...	...
$A(w_m, w_n)$	...	$\text{sim}(A(w_m, w_n), A(w_i, w_j))$	...
...	...	...	...

These meaning compositions correspond to the set of sub-propositions or RB-units that are derived from the word embeddings. From this similarity matrix at the RB-level, we can then compute the most activated sub-propositions. The final step is to extract the missing word from the most similar RB-unit. For example, if $A(w_i, w_j)$ is most similar with $A(w_m, w_n)$, we can then retrieve the word w_n from the composition to get the predicted word at the RB-level.

4.2.5 P-units

We also have to consider how the propositional units, or P-units, themselves give rise to a prediction about the upcoming word. In this case, we start from an incomplete proposition, e.g., *bigger(cup,?)*, and we compute the complete proposition that best fits this context. The idea is to measure the similarity between an

incomplete proposition like *bigger(cup,?)* and the set of all the complete propositions to find the most similar complete proposition and then derive the prediction from it. This derivation is very similar to what we described in terms of plausibility in Chapter 2. The idea is to find the continuation that would make the proposition the most plausible, given the incomplete information we possess.

To do this, we derive a representation for the incomplete proposition and a representation for the complete one and then measure the similarity between the two. We already mentioned that in both DORA and LISA, a composition of words is derived by adding vectors. This additive operation between a predicate and a filler is represented at the RB-level in 4.16, and I already presented how to measure the similarity between two sub-propositions in (85).

Here, it is slightly different because we are dealing with the similarity between an incomplete proposition and a complete one. The simplest way to derive those representations is to add all the word vectors together, but we have to keep in mind that we still have to consider a weighting parameter of 0.5 to differentiate between the contribution from the predicate and the fillers.

The representation for a proposition is the sum of two sub-propositions that are contained within it, i.e., the weighted sum of the two RB-units as shown in (87-a). Additionally, we have to do the same for the incomplete proposition, which amounts to the computation without the last word as in (87-b).

(87) a. bigger(cup,ball) = (larger+cup) + (smaller+ball) = larger + 0.5 cup + smaller + 0.5 ball

b. bigger(cup,?) = (larger+cup) + (smaller+?) = larger + 0.5 cup + smaller

Under this approach, we can measure the similarity between two entities that have a different number of words as illustrated in (88).

(88) $cos(P(w_1, w_2, w_3), P(w_5, w_6, w_7, w_8)) = cos(w_1{+}w_2{+}w_3, w_5{+}w_6{+}w_7{+}w_8)$

We then compute the similarity matrix at the P-level that gives the similarity measures between every incomplete propositions $P(w_1, w_2, w_3)$ and every complete ones $P(w_5, w_6, w_7, w_8)$ as illustrated in Table 4.4. We can find the most activated complete propositions with this matrix when a specific incomplete proposition is expressed. Going back to our example, this means that we are comparing *bigger(cup,?)* with every other possible complete proposition to determine the one complete proposition that is most similar with *bigger(cup,?)*. The dimension of the similarity matrix at the P-level if $N^3 \times N^4$, where the number of rows corresponds to the number of RB-units times the number of word embeddings, i.e., $N^2 \times N = N^3$, and the number of columns correspond to the number of combination between two RB-units, i.e., $N^2 x N^2 = N^4$.

Table 4.4 Similarity Matrix for every combination of RB-units

	...	$P(w_i, w_j, w_k, w_l)$	...
...	...	...	...
$P(w_m, w_n, w_o)$	...	$\text{sim}(P(w_m, w_n, w_o), P(w_i, w_j, w_k, w_l))$	...
...	...	...	...

Once we have determined the most activated propositions, we can extract the last word from these propositions, i.e., w_8 or w_l, like what we did for the RB-level, and use it as a lexical prediction at the P-level.

4.3 Computing a Prediction

The predictive model's basic premise is that the conditional probability of a word given a particular context, namely the predictability, is derived by considering different levels of representations. Each of these levels contributes to the derivation of a linguistic prediction as in Figure 4.18 (Doumas et al., 2018, adapted from their Figure 3).

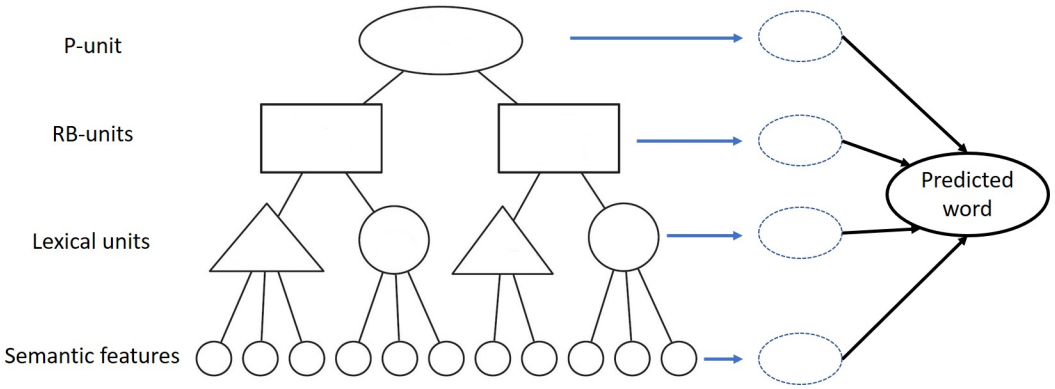


Figure 4.18 Four contributions to the derivation of a linguistic prediction

The contributions for the four representational levels we discussed, i.e., the semantic feature level, the lexical level, the RB-level, and the P-level, are all computed using activation-based semantic networks. This makes it easy to take all of them into account at once. Every level's output is a graph-like structure containing words that are all activated at different degrees. The idea is to simply add all the activations together to come up with the highest activated words. In other words, the final linguistic prediction is derived by summing up the activation of the linguistic prediction at each of these levels as in (89), where MAP is a set of most activated predictions at a given representational level.

(89) Prediction = $MAP_{\text{s.f.}} + MAP_{\text{PO}} + MAP_{\text{RB}} + MAP_{\text{P}}$

Figure 4.19 summarizes the derivation of the MAP for each representational levels. In Figure 4.19, the orange arrows represent the extraction of the information that leads to representing the four compositional units: P-unit, RB-unit, lexical unit, and semantic features. This operation is performed concerning the respective definition I presented for these four levels. The blue arrows from the units to the MAP represent the derivation of the most activated prediction, and it is performed using the similarity matrices at the P-level, at the RB-level, and the lexical level. This last one being used at both the lexical level and at the level of the semantic features. Each MAP corresponds to a set of predicted words, and the black arrows represent their combination into a set of predicted words that is computed by adding the four $MAPs$ together.

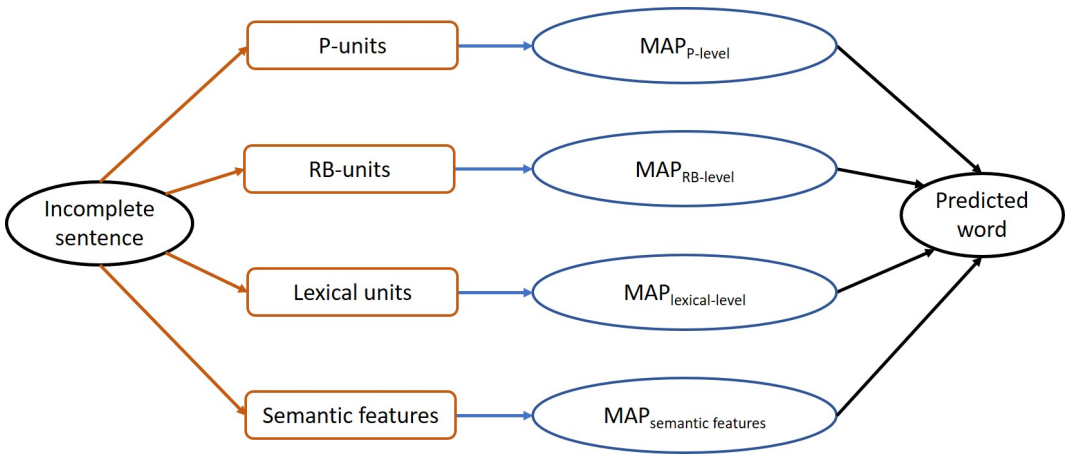


Figure 4.19 Depiction of the linguistic prediction process

It is important to note that the word with the highest activation level in total does not need to be the best candidate at each representational level. To illustrate the relative contribution from two levels of contributions, we can go back to our

example about the angry lion. In (27-a), reprinted here as (90), the lexical level might activate a particular subset of the lexicon, and *machete* may be the strongest candidate. However, the similarity features level for "sharp pointy object" might activate another subset of the lexicon, which would compete with the former so that *machete* might not have the highest cloze score, in the end, (Roland et al., 2012).

(90) The soldier jabbed the angry lion with ...

Another critical point is that in this basic model, the different representational levels cannot interfere with each other. They can be summed up, but they do not influence the activation level of one another. For example, the P-level is entirely independent of the semantic feature level regarding the activation-based semantic networks, not in terms of the representational units themselves. This means that we can simply add the contributions without modeling the possible interactions between two representational levels. This independence is mainly because we are not considering incrementality, so the model acts as if all the contributions are already derived when the summation is performed.[86]

Finally, it is important to note that because this model is activation-based, the most activated word should not be thought of as a 100% sure prediction, but it is instead an indication of the predictability of this word (Ferreira & Lowder, 2016). The output of this model is not linked with a particular realization of the close task, but it provides us a list of words categorized by their activation level concerning a particular context, i.e., $P(w_i|Context)$.

---

[86]This simplification is not always possible when using incremental models of linguistic prediction.

What about Syntax? In this chapter, I have developed a model of linguistic prediction following the empirical results discussed in Chapter 2, which supported the view that the semantic stream had precedence over the syntactic stream.[87] For example, the way we dealt with compositional units was by focussing on meaning compositions. However, even though we did not handle syntax explicitly, other than by eliminating by hand syntactic continuations that would be ungrammatical, it is partly taken into account as part of the different representational levels, since *John ate the apple* does not involve the same hierarchical representations as *The apple ate John*. One way to consider syntactic information without modifying the basic structure of the model presented here would be to integrate this information by modifying word embeddings training algorithms to account for syntactic information or by training these embeddings on parsed corpora (Linzen, 2019).

Instead, this thesis has taken a meaning-driven predictive approach since distributional representations, and activation-based semantic networks are primarily related to the semantic stream. When it comes to linguistic prediction, the meaning expressed at a different representational levels is important, but to help constrain the possible outcomes, we need to use syntactic information from the truncated sentence. As discussed in Chapter 3, Combinatory Categorial Grammar (CCG) (Steedman, 1996, 1999) fulfilled the three conditions of the desiderata for modeling linguistic prediction, and it is an excellent way to model the constraints imposed by the syntax. For example, in (91), if we are using the CCG terminology, the missing word category should be 'N' because it should combine with *to fly a* to form an 'S/NP' that could, in turn, combine with the rest of the sentence to give rise to a complete sentence of type 'S.'

(91) _____

[87] See also Baggio (2018) for a description of the relationship between these two processing streams.

The day was breezy so the boy went to the park to fly a ....

This hypothesis about the syntactic type of the missing word is obtained using the syntactic stream that uses morpho-syntactic rules to compute the missing category of words so that the complete sentence is compositional, i.e., that the sentence has the type 'S.' This syntactic combinatorial stream is thus capable of predicting the missing word category by computing the conditional probability of this syntactic type given the syntactic type of the rest of the sentence.

In this case, the conditional probability over the syntactic type is obtained by looking at all the possible continuations and picking the most probable one. In the syntactic stream, this operation is made possible because the parser is incremental, and so at every step of the process, the system automatically looks at potential syntactic types to complete the composition of the sentence.

In the present model, when computing a linguistic prediction, syntactic constraints limit the possibilities to ease the computation. At the end of the process, instead of measuring the similarity between all the word embeddings at the lexical level, we select only the word embeddings that correspond to the expected syntactic category. For example, if the sentence structure demands the linguistic prediction to be a noun, only the nouns will be selected as potential predictions. In the model of linguistic prediction presented in this thesis, the constraining influence originating from the syntactic processing is implicitly taken into account, even though it is not explicitly implemented in the model.[88]

---

[88]Even if syntax has a role to play in the process, it is essential to distinguish a meaning-driven process where syntax is only taken into account at the end of the derivation process, i.e. it is only after all the continuations have been derived that the syntactic constraints are applied, with a syntax-first process where these constraints would be used right at the beginning and would forbid some continuations to be activated. The way syntax is taken into account here is similar to what was described by Baggio (2018, p.186) as phrase structure level constaints (pslc) that may be imposed on the input string S (see also Section 2.4 and Section 5.5).

4.4 Worked-out Examples

This section illustrates how this linguistic prediction model works using three different kinds of truncated sentences: a high constraining sentence, a low constraining sentence, and a sentence where the influence of the previous context is crucial. Similarity measures were computed from the Gensim Python library using pre-trained word2vec-google-news-300 embeddings.[89],[90]

4.4.1 Example 1: High Constraining Sentence (HCS)

In a high constraining sentence, the number of possible continuations is low, and one continuation usually has a particularly high predictability score. For the sentence given in (92), it is the word *neck* that has the highest predictability as measured by Bloom & Fischler (1980) (96%) and by Block & Baldwin (2010) (97%).

(92) He loosened the tie around his ...

To compute the linguistic prediction, the first thing to do is represent the truncated sentence in terms of the four representational levels. This representation

---

[89] To avoid any prediction of very rare words, only the 100'000 most frequent words were used for the calculations.

[90] In this thesis, I use pre-trained word2vec word embeddings to show how to implement a vanilla model of linguistic prediction. In the future, it would be interesting to compare the results not only for different models of word embeddings but also for different pre-trained embeddings of the same model.

is illustrated in Figure 4.20. From there, we can derive the Most Activated Predictions (MAP) at each of these levels and combine them to obtain the linguistic prediction.

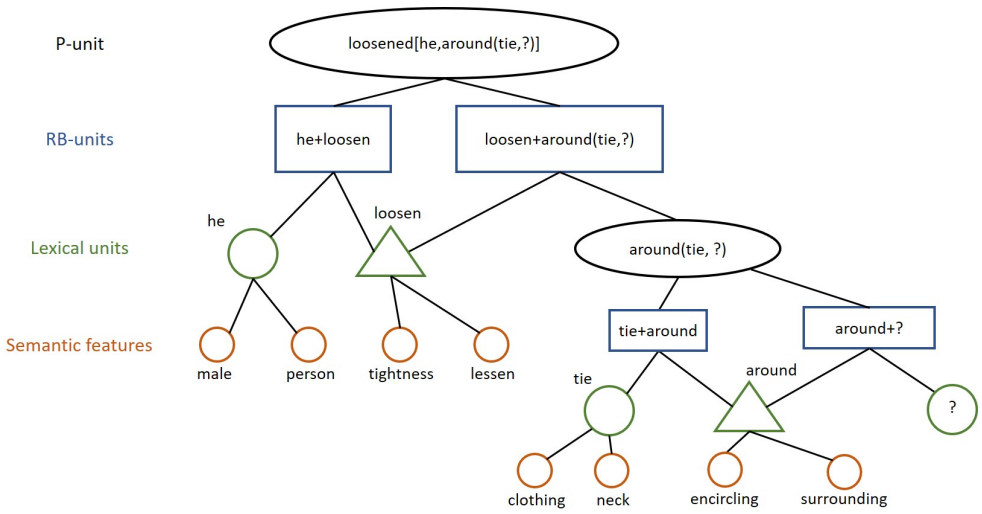


Figure 4.20 Representation of the proposition *loosened(he,around(tie,?)*

The representation depicted in Figure 4.20 is different from the representations we could expect in LISA and DORA, and this is due to the fact that I am using word embeddings to compute similarities, and the words or the expressions depicted in the representation must be present as word embeddings in the corpus. For example, instead of *loosener+he* and *loosens+around(tie,?)*, I used *loosen* for the RB-units because both *loosener* and *loosens* are not in the 100'000 most frequent words. However, to remain true to the principle behind DORA and LISA, *loosen* is taken into account twice by the model because it is linked with two different RB-units.

4.4.1.1 MAP at the Lexical-level: HCS

At the lexical unit level, the MAPs are computed by finding the words that are the most similar to the words expressed in the first part of the sentence. Using the similarity matrix that we have derived at the lexical level, we measure the similarity between the predicted words and the lexical units present in (92).

(93) MAP_L = Lexical-unit that is the most similar to *he*, *loosen*, *loosen*, *tie*, *around*.

Table 4.5 First five MAP for the Lexical-units of (92)

win	3.285
back	3.020
tie	2.983
semifinal	2.364
finish	2.176

In Table 4.5, the activation level for every MAP is the sum of the contributions from the five lexical units given in (93). As explained before, only the words associated with the relevant grammatical categories are selected in this model, and the rest are discarded. In this Table and all subsequent tables, we only represent the words with the relevant grammatical categories. In the case of (92), the predicted word should be a noun. Interestingly, at the lexical level, the word *neck* is not present at all in the MAPs, which means that it is not activated.

4.4.1.2 MAP at the Semantic Feature-level: HCS

At the Semantic Features (SF) level, the MAPs are computed similarly to that we have done for the lexical level, but here the MAP corresponds to the most similar words to the semantic features activated by the lexical units. Using the same similarity matrix for the lexical units, we measure the similarity between the predicted words and the semantic feature units present in (92).

(94) MAP_{SF} = Lexical-unit that is the most similar to *male*, *person*, *tightness*, *lessen*, *clothing*, *neck*, *encircling*, *surrounding*.

Table 4.6 First five MAP for the SF-units of (92)

forearm	8.560
thigh	7.626
tendinitis	7.509
rib_cage	6.875
groin	6.863

At the level of the semantic features, the word *neck* does not appear in the first five MAPs, but it is activated. Its activation is 5.071.[91]

[91] The difference in the absolute values of the activation for the MAPs in Table 4.6 and those in Table 4.5 arises because the normalization factor is not the same for the different levels of similarity. When combining the contribution from two different levels, we have to renormalize the activation so that every level contributes equally.

4.4.1.3 MAP at the RB-level: HCS

At the level of the RB-unit, the MAPs are computed by finding the complete RB-units that are the most similar to the incomplete RB-unit $around_{RB}$. Using the similarity matrix that we already derived for the RB-units, we measure the similarity between the complete RB-units and the incomplete RB-unit, which consists, in the case of (92), of the lexical unit *around*. Once we have computed the MAP at the RB-level, we can retrieve the last word of the RB-unit, which corresponds to the prediction of a lexical unit.

(95) MAP_{RB} = RB-unit that is the most similar to $around_{L}$

Table 4.7 First five MAP for the RB-units of (92)

around+*back*	0.934
around+*past*	0.922
around+*neck*	0.842
around+*tie*	0.8412
around+*legs*	0.836

In Table 4.7, the words in italic correspond to lexical prediction retrieved from the RB-units.

4.4.1.4 MAP at the P-level: HCS

At the level of the P-unit, the MAPs are computed by finding the complete propositions that are the most similar to the incomplete proposition given in (92). Using the similarity matrix we derived for the complete propositions, we measure the similarity between these complete propositions and the incomplete proposition, which consists, in the case of (92), of *loosen(he(around(tie,?))*. Once we have computed the MAP at the P-level, we can retrieve the last word of the proposition corresponding to the prediction at the lexical level.

(96) MAP_P = Proposition that is most similar to $loosen{+}he_{RB}{+}tie{+}around_{RB}{+}around_L$

Table 4.8 First five MAP for the P-unit of (92)

(loosen+he) + (tie+around) + (around+*back*)	0.983
(loosen+he) + (tie+around) + (around+*neck*)	0.963
(loosen+he) + (tie+around) + (around+*legs*)	0.960
(loosen+he) + (tie+around) + (around+*knees*)	0.958
(loosen+he) + (tie+around) + (around+*tie*)	0.952

In Table 4.8, the words in italic corresponds to lexical prediction retrieved from the propositions.

4.4.1.5 Deriving the Linguistic Prediction: HCS

To derive the linguistic prediction, we have to combine the MAPs obtained at every level by summing up the activation of the linguistic prediction as in (97). In order to treat the contributions equally, we normalize the MAP at every level in terms of their relative activations with respect to the number of possible predictions at that level.

(97) $\text{Prediction} = MAP_{\text{SF}} + MAP_{\text{PO}} + MAP_{\text{RB}} + MAP_{\text{P}}$

Table 4.9 First five MAP for the sentence (92)

back	0.055
neck	0.052
thigh	0.051
rib_cage	0.051
legs	0.051

To derive the final MAP at every level, I have summed up the activations of the different realizations of the words, e.g., the contribution from the words *Hand*, *hands*, *hand*. On the other hand, the contribution from related words that have different grammatical types was not included in this sum, e.g., *handy* or *handling*. There are no clear indications in Bloom & Fischler (1980) or Block & Baldwin (2010) as to whether plurals were recorded as different answers, but I chose to combine these contributions to reflect the fact that this is a comprehension-centric model of linguistic prediction where it is the meaning of a word that is predicted instead of a specific lexical entry. In Table 4.9, I assumed that every level is

contributing equally to the derivation of the linguistic prediction.[92] Even though the relative contribution from the different levels might be difficult to grasp from these tables, we can still readily see every levels contains at least one continuation that is present in the overall prediction. The word *thigh* for example is only present in the five most activated prediction at the semantic features level, while the lexical level contains four candidate words that are expelled from the top five when combined with the other levels.

4.4.2 Example 2: Low Constraining Sentence (LCS)

In a low constraining sentence, the number of possible continuations is significant and there are no continuations that have particularly high predictability score. For the sentence given in (98), the cloze scores are the following (Bloom & Fischler, 1980): *hands* (49%) , *glove* (32%), *mitt* (8%), *teeth* (4%).

(98) Dan caught the ball with his ...

To compute the linguistic prediction, the first thing to do is represent the truncated sentence in terms of the four representational levels. This representation is illustrated in Figure 4.21. From there, we can derive the Most Activated Predictions (MAP) at each of these levels and combine them to obtain the linguistic prediction.

---

[92] In this model, it is possible to modify the contribution weight of a specific level. This issue is discussed in Chapter 6.

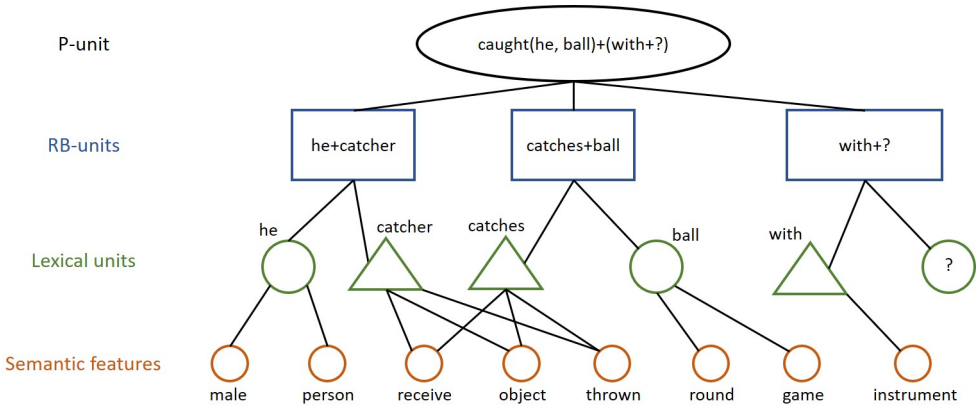


Figure 4.21 Representation of the proposition *caught(he,ball)+with(?)*

4.4.2.1 MAP at the Lexical-level: LCS

Using the similarity matrix that we derived at the lexical level, we measure the similarity between the predicted words and the lexical units present in (98).

(99) MAP_L = Lexical-unit that is the most similar to *catcher*, *he*, *catches*, *ball*, *with*.

Table 4.10 First five MAP for the Lexical-units of (98)

catcher	7.476
quarterback	7.038
shortstop	6.825
pitcher	6.458
baseman	5.778

At the lexical level, the word *hands* does not appear in the first five MAP, as its

activation is only 0.161. The activation of the word *glove* is 0.688.

4.4.2.2 MAP at the Semantic Feature-level: LCS

Using the same similarity matrix for the lexical units, we measure the similarity between the predicted words and the semantic feature units present in (98).

(100) MAP_{SF} = Lexical-unit that is the most similar to *male*, *person*, *receive*, *object*, *thrown*, *round*, *game*, *instrument*.

Table 4.11 First five MAP for the SF-units of (98)

game	6.462
semifinals	5.465
quarterfinals	5.453
tournament	5.400
finals	4.668

At the level of the semantic features, the word *hand* is not activated, but the word *glove* is. Its activation is 0.100.

4.4.2.3 MAP at the RB-level: LCS

Using the similarity matrix that we already derived for the RB-units, we measure the similarity between the complete RB-units and the incomplete RB-unit, which

consists, in the case of (98), of the lexical unit *with*. Once we have computed the MAP at the RB-level, we can retrieve the last word of the RB-unit, which corresponds to the prediction at the lexical level.

(101) MAP_{RB} = RB-unit that is the most similar to $with_L$

Table 4.12 First five MAP for the RB-units of (98)

with+*hands*	3.275
with+*glove*	2.124
with+*socks*	0.760
with+*toes*	0.738
with+*legs*	0.734

In Table 4.12, the words in italic correspond to lexical prediction retrieved from the RB-units.

4.4.2.4 MAP at the P-level: LCS

Using the similarity matrix that we already derived for the complete propositions, we measure the similarity between the complete propositions and the incomplete proposition, which consists, in the case of (92), of the RB-units *catcher+he*, *catch+ball*, and of the lexical unit *with*. Once we have computed the MAP at the P-level, we can retrieve the last word of the proposition, which corresponds to the prediction at the lexical level.

(102) MAP_P = Proposition that is most similar to $he{+}catcher_{RB}{+}catches{+}ball_{RB}{+}with_L$

Table 4.13 First five MAP for the P-unit of (98)

(catcher+he) + (catches+ball) + (with+*hands*)	3.821
(catcher+he) + (catches+ball) + (with+*glove*)	1.882
(catcher+he) + (catches+ball) + (with+*mitt*)	1.876
(catcher+he) + (catches+ball) + (with+*fingers*)	1.855
(catcher+he) + (catches+ball) + (with+*socks*)	0.944

In Table 4.13, the words in italic correspond to lexical prediction retrieved from the propositions.

4.4.2.5 Deriving the Linguistic Prediction: LCS

To derive the linguistic prediction, we have to combine the MAPs obtained at every level. This linguistic prediction is derived by summing up the activation of the linguistic prediction at each of these levels as in (103), but we first have to normalize the MAP at every level so that they each contribute to the same proportion to the linguistic prediction.

(103) $\text{Prediction} = MAP_{\text{SF}} + MAP_{\text{PO}} + MAP_{\text{RB}} + MAP_{\text{P}}$

In 4.14, we see the order of the first three MAPs is the same as the one obtained by (Bloom & Fischler, 1980). However, the word *teeth* is not activated by the model. When comparing the contribution coming from the four levels, we readily see that

Table 4.14 First five MAP for the sentence (98)

hands	0.164
glove	0.118
mitt	0.078
fingers	0.077
socks	0.040

here, both the top five continuation at the lexical and the semantic features levels seems to contribution less to the final combined prediction. This is not to the first five MAP for the sentence (98) are not activated at these levels, but only they stand out less than it was the case for the HCS example. This should not be surprising at this latter case is less constrained than the former.

4.4.3 Example 3: The influence of the context (IoC)

In some cases, the prior context might increase the probability of a particular non-conventional combination of words. For example, Nieuwland & Van Berkum (2006) showed that, when given a suitable discourse context, participants processed animacy-violating predicates more easily, i.e. *the peanut was in love*, than canonical predicates like *the peanut was salted*. Their result was obtained by comparing the measured N400 effect with and without a context that supported this animacy-violation. In the example (104) taken from Nieuwland & Van Berkum (2006), the association between the meaning of *peanut* and the possible continuation is modified by the use of a specific context where the peanut is attributed

anthropomorphic characteristics.[93]

(104) [Full context] A woman saw a dancing peanut who had a big smile on his face. The peanut was singing about a girl he had just met. And judging from the song, the peanut was totally crazy about her. The woman thought it was really cute to see the peanut singing and dancing like that. The peanut was [salted/in love], and by the sound of it, this was definitely mutual. He was seeing a little almond.

In this section, I use my model to derive the linguistic prediction for the sentence in (105). The goal here is to test whether it could tackle such particular cases where a previous context influences the expectations regarding the possible continuations. In this chapter, I am only considering the truncated sentence, and in Chapter 5, I will take into account the previous context.

(105) The peanut was ...

In the previous examples, we derived a prediction for a specific argument of a predicate, but here we have to predict a predicate directly. This difference reverberates in the way the incomplete proposition is represented in Figure 4.22. Here, the RB-level and the P-level are merged because we only have a single argument predicate which means we can omit the P-level for the derivation. From this representation, we can derive the Most Activated Predictions (MAP) at each of these

[93] Another similar result was obtained by Cosentino et al. (2017) when they compared the N400 components for cases with Telic and Atelic noun-verb combinations. They considered two kinds of discourses: a neutral context, and a context that induced a new function for an object. Their results showed that the prior function-inducing context was reducing significantly the N400 component for the non-Telic noun-verb combinations cases.

levels and combine them to obtain the linguistic prediction.

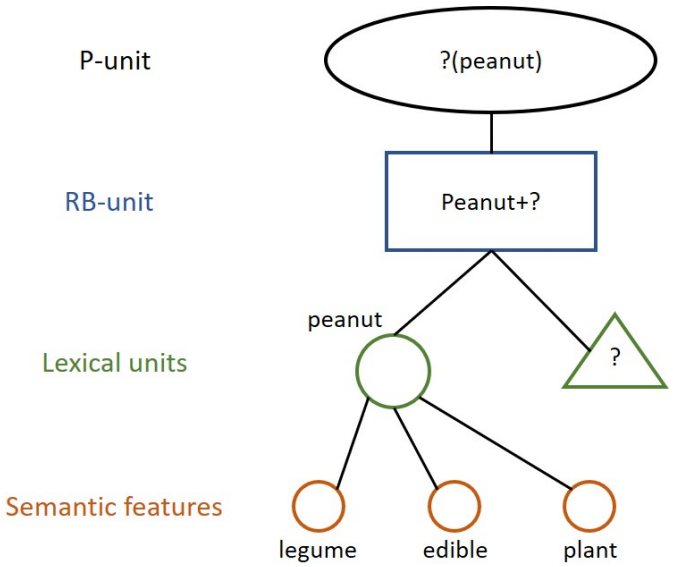


Figure 4.22 Representation of the proposition *?(peanut)*.

In the case at hand, the goal is not to predict the upcoming word, but to illustrate the different expectations regarding the processing of the words *salted* and *love*.[94]

4.4.3.1 MAP at the Lexical-level: IoC

Using the similarity matrix that we derived at the lexical level, we measure the similarity between the predicted words and the lexical units present in Figure 4.22.

(106) MAP_L = Lexical-unit that is the most similar to *peanut*

[94]In the wiki-news corpus, the expression *in-love* is present but is very rare, i.e., its rank is 664'236. In this calculation, I instead use the word embedding *love*.

Table 4.15 First five MAP for the Lexical-unit of (105)

dairy	2.685
milk	2.567
meat	2.474
pork	2.290
rice	2.263

4.4.3.2 MAP at the Semantic Feature-level: IoC

Using the same similarity matrix for the lexical units, we measure the similarity between the predicted words and the semantic feature units present in Figure 4.22.

(107) MAP_{SF} = Lexical-unit that is the most similar to *seed*, *edible*, *plant*.

Table 4.16 First five MAP for the SF-units of (105)

plant	4.113
vegetables	4.098
wheat	3.612
corn	3.277
rice	2.984

4.4.3.3 MAP at the RB-level: IoC

Using the similarity matrix that we derived for the RB-units, we measure the similarity between the complete RB-units and the incomplete RB-unit, which consists, in the case of Figure 4.22, of the lexical unit *peanut*.

(108) MAP_{RB} = RB-unit that is the most similar to $peanut_L$

Once we have computed the MAP at the RB-level, we can retrieve the last word of the RB-unit, which corresponds to the prediction at the lexical level.

Table 4.17 First five MAP for the RB-unit of (105)

peanut+*potato*	0.960
peanut+*vegetable*	0.960
peanut+*peach*	0.960
peanut+*tomato*	0.955
peanut+*strawberrry*	0.954

In Table 4.17, the words in italic correspond to lexical prediction retrieved from the RB-units.

4.4.3.4 Deriving the Linguistic Prediction: IoC

To derive the linguistic prediction, we have to combine the MAPs obtained at every level. This linguistic prediction is derived by summing up the activation

of the linguistic prediction at each of these levels as in (109), but we first have to normalize the MAP at every level so that they each contribute to the same proportion to the linguistic prediction.

(109) $\text{Prediction} = MAP_{\text{SF}} + MAP_{\text{PO}} + MAP_{\text{RB}} + MAP_{\text{P}}$

Table 4.18 First five MAP for the sentence (105)

milk	0.028
wheat	0.028
dairy	0.027
meat	0.027
grain	0.027

We see from Table 4.18 that the linguistic prediction does not make much sense, but, as described before, the goal here was not to predict the upcoming predicate but to compare the activation of *love* and *salted*. As it turns out, neither of them is activated in the model. However, we do have activation for the word *salt*, which is semantically related to *salted*. Its activation is 0.007. Contrary to what we had before, the contribution coming from the RB-level seems to be minimal here as the first five MAP for the sentence (105) looks very similar to the most activated continuations for the lexical and the semantic features levels.

In this calculation, the prior context is not taken into account, and the information present in this prior context might influence the relative activation between the word *love* and *salted*. In Chapter 5, we revisit this example while taking into account the prior context to see how it affects the relative activation between the two.

4.5 Summary

In this chapter, I have presented a model of linguistic prediction based on the pre-activation of words at different levels of representations. The basic idea behind this approach is that linguistic prediction can be modeled as a series of activation paths within the conceptual space. Every time a meaningful linguistic unit is processed, it lights up a corresponding region in conceptual space, and from this region, the activation spreads to conceptual neighbors similar to this linguistic unit. In this model, I have considered four levels of spreading activation, and each of them is linked with a particular level of linguistic representation. This multi-level view of linguistic prediction is in line with what we described in Chapter 2 about the parallel architecture of Baggio (2018).

Similarity, Predictability, and Plausibility The predictability is the cloze values of a word, and it is equivalent to the conditional probability of using a word given a particular context. To derive the result of this conditional probability, we considered four representational levels and their associated activation-based semantic networks. These networks are constructed from the similarity measures between the different linguistic units, i.e., semantic features, lexical units, RB-units, and P-units.

Usually, semantic similarity only refers to the similarity at the lexical level, but here, the similarity is used to derive the most activated predictions at different representational levels. When measuring the similarity between linguistic units, we have to take into account the nature of these linguistic units and how they are formed. Consequently, the way this similarity is computed varies for each representational level. Going back to the distinction we underlined in Chapter 2

between similarity and plausibility, we can associate them to specific representational levels.

For example, the linguistic prediction process involving the notion of plausibility is tantamount to the similarity measured at the P-level, because the linguistic prediction is derived using the correspondence between the predicted word and the whole sentence. Similarly, the passive co-activation process we discussed in Chapter 2 would correspond to the similarity at the lexical level within this multi-representational model of linguistic prediction.

In summary, in this model, the predictability is obtained using different representational levels of MAP, and it encompasses what was previously described as passive co-activation and plausibility. This view about linguistic prediction helps disambiguate these three notions and explains why it is difficult to treat them separately.

In the last section of this chapter, I have presented three examples to illustrate could be implemented. In the case of the High Constraining Sentence, the model was able to retrieve the word *neck* as the second most activated prediction, just behind the word *back*. When I considered the Low Constraining case, the ordering for activations for the words *hands*, *glove*, and *mitt* exactly corresponds to the ordering for the cloze scores obtains by Bloom & Fischler (1980). Put together, these results supports the empirical adequacy of this model. Finally, in the case where we had to compute the relative activation of the word *love* and the word *salted*, the model did not activate the word *love* and it only activated *salt* instead of *salted*. In the following chapter, I come back to these worked-out examples, and I integrate the contribution coming from the coordinated aspect of language.

CHAPTER V

LINGUISTIC PREDICTION AND COORDINATION

5.1 Communication and Cloze task

The previous chapter presented how a prediction for the upcoming word can be derived from the meaning composition represented using different linguistic units, namely the semantic features level, the lexical level, the RB-level, and the P-level. However, when we add these contributions together to derive a linguistic prediction, we are not considering the coordination aspect of communication.

As I already described in Chapter 3, the meaning composition is about the meaning that emerges from combining the words presented in a sentence. When dealing with coordination, we also have to take into account the speaker that uttered this sentence because communication involves two agents, and each of them has an active role to play during a linguistic exchange.

The present chapter discusses this coordination aspect by considering that sentences, and more generally linguistic information, be thought of as communicative acts. This perspective about coordination has significant consequences for how a linguistic prediction is derived.

5.1.1 Communication and Coordination

Communication originates from an interaction between two agents: a speaker conveys a particular meaning and produces a sentence, while a hearer retrieves the conveyed meaning that was intended. One important characteristic of communicative interaction is the inherent uncertainty that is involved. This uncertainty arises naturally when the hearer interprets as language itself is uncertain or ambiguous.[95]

For example, in (110-a), the meaning of the word *bank* is uncertain because it could either refer to a financial institution or a rising ground bordering a river. This is just one example of lexical ambiguity, but we also have other kinds of uncertainties. In (110-b), we have morphological uncertainty because the *-s* could have different meanings, and we have a phonological uncertainty in (110-c) because the same phonemes are attached to different meanings. Finally, in (110-d), there is uncertainty concerning the referent of the turtle. In a case where the only information we have about the turtle is that Moira took it, then we might be able to grasp a generic meaning of a turtle, but we would not be able to attribute this act of being taken to a specific turtle.

(110) a. The fisherman went to the bank.

b. *Mark-'s*, *pen-s*, *eat-s*

c. *too*, *two*, *to*.

---

[95]Piantadosi et al. (2012) and Gibson et al. (2019) used the term *ambiguity* to refer to lexical and syntactic ambiguity. In this thesis, I use *uncertainty* instead of *ambiguity* because *uncertainty* could also encompass other notions such as *vagueness*, *underdetermination*, and *context sensitivity* which are all present in communication. See Sennet (2016) for a detailed definition of these notions.

d. Moira took the turtle.

These examples show that linguistic units by themselves are not informationally strong enough to compensate for their innate meaning flexibility. However, this natural uncertainty is not a problem for communication per se because we seem to have relatively few problems understanding each other during a day-to-day conversation. Most importantly, having to deal with uncertainty should not be thought of as incompatible with the view that language is a sound communication system. In a perfect communication system, i.e., a maximally precise communication system, uncertainty would be non-existent, and each word, or each linguistic unit, would be expressing only one meaning and vice-versa. The correspondence between meaning and linguistic form would thus be perfectly bi-directional.

A perfect communication system has some advantages because listing unique mappings between words and meanings would decrease lexical uncertainty. However, this implies that the number of mappings we have to learn would dramatically increase. Additionally, in this perfectly non-ambiguous communication system, to convey some simple meaning, we would have to use perfectly specific words that would only describe a very particular meaning and nothing else.

Instead of learning new words that are more specific and less ambiguous, another option would be to use more words to avoid uncertainty, as in (111). However, this method is not very efficient because it increases the complexity of the production phase tremendously (Piantadosi et al., 2012). For example, in (111) we have to produce at least three more words to specify the meaning of the single word *bank*.

(111) a. The fisherman went to the bank.

b. The fisherman went to the bank of the river.

c. The fisherman went to the bank to make a deposit.

In fact, in a real-life conversation, a speaker would most probably choose to utter (111-a) and would rely on the hearer to be able to grasp the intended meaning of *bank*. Instead of spending extra energy on trying to be as specific as possible, the speaker puts the burden of interpretation on the hearer's shoulder, which has to disambiguate the message. In everyday conversation, the speaker almost always takes this effortless path, and, interestingly, we still rarely face communicative complications because of this decision. Language is not a maximally precise communication system, but it seems that maximum precision is not compulsory for successful communication.

The main reason why ambiguous language communication does not have to be perfect is that hearers are very good at disambiguating meanings in context (Piantadosi et al., 2012). As illustrated in (111), the hearer is usually able to grasp the meaning of (111-a) from the information present in the context of production, and from this perspective, it would have been a wasted effort for the speaker to be more specific.

The key is that a hearer can disambiguate using information from the context, and it is this contribution from the context that makes language communication efficient despite its inherent ambiguity (Gibson et al., 2019). This perspective about language is in line with the description given in Clark (1996) where language use is defined as error-prone and more context dependant than what is often suspected. The idea that an interpreter uses the context to disambiguate the meaning of a word or a sentence is a primordial feature of a communication system because it improves efficiency both on the production and the interpretation side.

Assuming the context is disambiguating, we no longer need a perfectly efficient

communication system because it is cognitively easier and still communicatively successful to produce ambiguous sentences (Piantadosi et al., 2012). Furthermore, this means that we can re-use the same words and expressions to refer to different things. It also implies that a speaker does not have to be maximally precise when producing a sentence because the hearer can complete the meaning of that sentence using the information from the context (Piantadosi et al., 2012).[96]

In a balanced communication system, in addition to the context, we also need words and sentences to guide the interpretation process because the context is usually not enough to disambiguate all complex meanings.[97] In other words, we have a contribution coming from the words, but we also have a contribution coming from the context. When it comes to linguistic prediction, there are no principled distinctions between the two (Casasanto & Lupyan, 2015), and they both have to be taken into account.

In Chapter 4, I discussed the contribution from the compositional aspect of communication, and I used the term *context* to refer to the local context, i.e., the linguistic information explicitly derived from the words and the compositions that are expressed in a sentence. This local context exists in addition to the global context that is derived beyond the single sentence. I use the term *global* here to avoid any confusion with the concept of *broader context* that was briefly discussed in Chapter 4. The distinction between *broader context* and *local context* is measured in terms of their distance with respect to the target linguistic unit.

---

[96]Although it was worded quite differently, this argument for the contribution from the context is somewhat parallel to what has been described as the 'Semantic Underdeterminacy view' in the literature about the distinction between semantic and pragmatics (Carston, 2002; Recanati, 2005; Bach, 2004).

[97]See Piantadosi et al. (2012) for a mathematical argument that the contextual contributions help minimize the entropy (the ambiguity) of a linguistic unit whenever the context is informative.

For example in (112), the terms *breezy* and *day* are considered to be part of the broader context, whereas *to fly* is part of the local context compared with the point of truncation.

(112) *The day was breezy, so the boy went to the park to fly....*

What brings the broader and the local context together is that they refer to the information expressed within a sentence. On the other hand, the term *global context* I use here refers to the sum of information available to the interpreter at the time of the prediction. This information could be derived from linguistic sources like the previous sentences or non-linguistic perceptual inputs originating from vision or hearing.[98]

Thus, the global context encompasses both the broader context and the local context because they also contribute to the linguistic prediction and any other forms of contextual information relevant to their understanding of a given situation at a specific time.[99] Following this definition, the distinction between *local context* and *global context* is related to the amount of information taken into account and not about the nature of this information. Others like Bach (1997, 2001) have separated the context into two kinds depending on the nature of the derivation of the contextual information: a narrow context corresponding to the semantic interpretation and a wider context corresponding to the pragmatic interpretation.[100]

---

[98]Other information like general knowledge about the world, or our own prejudices could also be included in the global context. However, for the purpose of this thesis, we only consider linguistic information from the preceding discourse.

[99]I coined the term *global context* to avoid any confusion with the many different definitions or uses of the term *context* in Linguistics or Cognitive Science.

[100]See Carston (2004); Recanati (2005, 2002) for more explanations about this distinction.

In this thesis, the global context and the local context are not distinguished in terms of their informational content, only in terms of their scope. Also, I do not differentiate between linguistic and non-linguistic contextual information. Finally, when it was argued that the context played a disambiguating role during communication, the kind of context that was referred to was the global context, i.e., the sum of the information that can serve as an informational foundation upon which the hearer can derive an interpretation or, in our case, a linguistic prediction.

5.1.2 Coordination and Context

The importance of this global context follows from the very nature of the coordination between two communicators.[101] To understand how vital is this relationship between context and coordination, we can go back to the expression presented in Chapter 3 about the nature of the anticipatory process. In (113), two kinds of coordination are linked with the interpretation and the production phase of a linguistic prediction.

(113)

$$\text{Word}_{\text{Pred}} \propto [f_1(\text{partial utterance,world}) \times \text{P(world)}]_{\text{Inter}} \times [f_2(\text{utterance,world})]_{\text{Prod}}$$

The first kind of coordination is represented by the function f_1 that maps a partial

[101] For the rest of this thesis, unless it is specified otherwise, I am using the term *context* to mean *global context*.

utterance into a state of the world, i.e., f(partial utterance,world). In Chapter 3, we called this the mapping functions between states of the world and utterances, and the hearer needed to know which mapping function was used by the speaker to avoid any mismatches like in the cinema example about Margot and Fiona. This mapping function is directional, and it goes from the partial utterance to the state of the world. For example, if a sentence is of the form presented in (114), the hearer has to update her representation of the state of the world to correspond with the new information that Viola has three something. Keeping in mind that the global context is the sum of all information available to the hearer at a time t, the global context corresponds to the world's state, and we could treat this mapping function as a function linking the global context and the partial utterance.[102]

(114) Viola has three ...

The second function term in (113), i.e., P(world), does not directly involve a coordination component because this term is about the correspondence between states of the world at two contiguous times. Put differently, it is similar to the transition probability for going from state to state without any external input, but because it is not related to coordination, I will not discuss it further at this point.[103]

---

[102]This mapping function between the global context and the information expressed at the word level could be transposed in scene recognition studies in terms of the relationship between the *gist of a scene* and the individual objects present in that scene. See Oliva & Torralba (2001); Oliva (2005); Torralba & Sinha (2001) for more details about perceptual studies.

[103]An example of these kinds of inferences would be to use information already known to derive a new conclusion, i.e., to auto-generate a new state of the world without processing any new inputs.

The other important mapping function from (113) is the one involved in the production phase of the prediction process, namely the mapping f_2 between the world state and the utterance. This mapping is about choosing utterances that can be used to describe the world in this particular situation. Here, the mapping function goes from the state of the world, i.e., the global context, to an utterance.

(113) thus illustrate the importance of the global context that naturally emerges from the coordination involved between the hearer and the speaker when considering linguistic prediction. There are two kinds of coordination: the first kind involves the speaker, and it maps his partial utterance with the context, and the second kind involves the hearer, and it maps the context to an utterance.

To better understand the different roles of these two kinds of coordination, I start from the traditional coding-decoding perspective of communication and adapt it to illustrate the global context's contribution. In the conduit metaphor view of linguistic communication (Reddy, 1979), the speaker intends to convey a particular meaning, and, after that, he encodes this meaning using linguistic units. The hearer has then to decode these linguistic units to retrieve the conveyed meaning that the speaker intended. The crucial contribution from the context is often not explicitly represented in this traditional coding-decoding view. To illustrate the contribution from the context, I use the recent Information-Theoretic picture of Communication presented in Gibson et al. (2019), and I present an updated version of their Figure 1 in Figure 5.1 where the context is represented explicitly.

In Figure 5.1 (Gibson et al., 2019, modified from their Figure 1), the context has to be accounted for by the speaker at the production phase and by the hearer at the interpretation phase. To produce the best possible utterance (the signal), the speaker has to determine the words that will, in conjunction with the contextual information at time t_i, be interpreted as intended. On the other hand, the hearer

must use this signal in conjunction with the context at time t_{i+1} to retrieve the intended conveyed meaning.

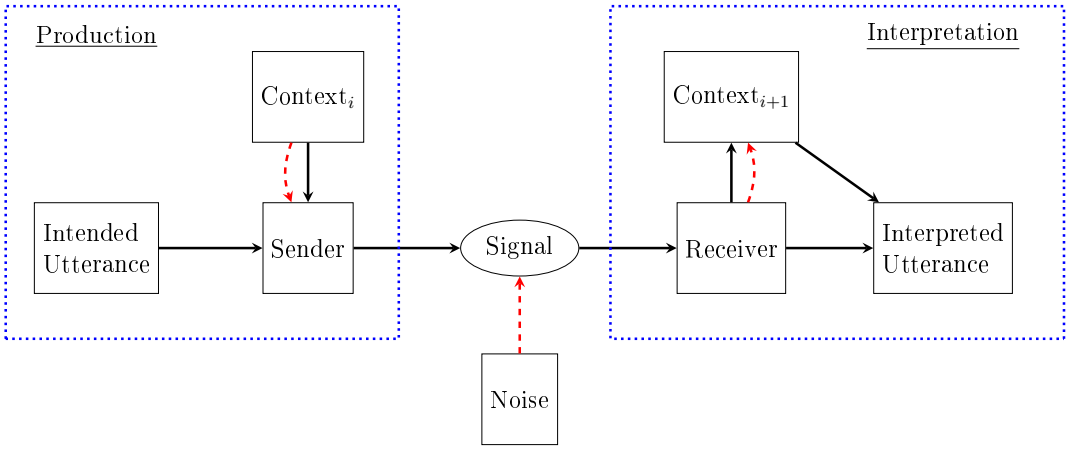


Figure 5.1 A schematic depiction of linguistic communication

Another essential feature of linguistic communication represented in this graph is the potential noise within the communication channel. In Chapter 6, I discuss the fact that the representation of the context is lossy (Futrell et al., 2020). This means that the representation of a word becomes noisy as time passes, which, in turn, implies that words that are close to the predicted target word should contribute more (Futrell et al., 2019, 2020). This lossy-context noise is related to the incremental decoding process, and it is part of the potential noise that could play a structural role during linguistic communication. Similarly, we can also have a lossy input which would be related to the perceptual uncertainty of the signal (Levy, 2011). For example, if in a crowded room or a stadium full of noises, it can be challenging to understand the uttered words. Having a noisy input is not mutually exclusive from having a lossy-context representation since the two originate from different cognitive mechanisms. The former being related to the correct representation of the word, while the former is instead related to

the lossy nature of its representation in memory.

The third kind of noise, illustrated in Figure 5.1 as a curved red arrow, originates from the extra-linguistic information. This noise is related to the uncertainty about the state of the world from the hearer's perspective and the uncertainty about the state of the world from the speaker's perspective, respectively. For example, we could be uncertain about the current state of the world, i.e., we might not be sure whether a given statement is true or not, or we might not be sure whether the speaker knows about another statement being true.

The presence of these noises/uncertainties is not problematic as they do not prevent us from communicating. When facing a noisy channel, a hearer uses what is known as rational statistical inferences (Gibson et al., 2013) to interpret the correct meaning despite these noisy representations, but I will come back to these inferences about uncertainty in a later section of this chapter.

In Summary, in Figure 5.1, we see that during the production phase, the coordination is between the context$_i$ and the intended utterance, and it involves the speaker, whereas, in the interpretation, the coordination is between the context$_{i+1}$ and interpreted utterance. It does not directly involve the speaker. This distinction is vital because this latter coordination is present independently of the presence of the speaker. In other words, no matter the origin of the signal, interpretation always involves coordinating a representation of the context with an interpreted utterance. These two kinds of coordination are essential for linguistic prediction, and I will discuss how to model their contributions shortly, but first, we still have to specify the role of this coordination aspect when it comes to the cloze task.

5.1.3 Context and Cloze Task

The main difference between full-fledge linguistic communication and performing a cloze task is that there is no strong relationship between the sentences during the latter because, in a cloze task, every sentence is situated in a context of its own. Consequently, it is more difficult to build a coherent global context from only one sentence.[104] However, this does not proscribe the existence of a contextual contribution because all single sentences are nonetheless contextually situated. As described by Bach (2004, p.39), when a hearer encounters sentences in isolation, "certain default assumptions are made about the circumstances of the utterance". These default assumptions or heuristics arise because language is about regularities (Gastaldi, 2020). This regularity is present at different linguistic levels, as we already discussed in Chapter 4, and it is also present beyond the utterance level.

When trying to predict the next word of a truncated sentence, our experience concerning linguistic communication shapes our expectations towards upcoming information (Brehm et al., 2019). During our lifetime, we not only learn to predict the next upcoming word from the previous linguistic context, but we also learn the correspondence between global context and upcoming information. To illustrate this correspondence between the context and the linguistic prediction, we could go back to the example discussed in Chapter 1.

(115) a. The kind old man asked us to ...

[104]Brothers et al. (2020) make a distinction between *Deep interpretation* when a hearer can derive a global context, and *Surface interpretation* when a hearer is not able to derive a global context. In Chapter 6, I come back to these different depths of interpretation and argue that they could play a role when modeling linguistic prediction.

b. [After a day-long trek, we arrive at a mountain hut, and we see that the other guests have already joined the host at the table.] The kind old man asked us to ...

With a truncated sentence like (115-a), it is very difficult to predict the next word, because the context is not constraining enough, and, as a result, we end up with many different continuations (Bloom & Fischler, 1980): *stay* (26%), *help* (21%), *leave* (10%), *dinner* (5%). However, if we add a global context before the sentence as in (115-b), then the cloze values of the possible continuations would certainly be different, i.e., *dinner* would most probably have a higher cloze score. The idea here is to argue that without any other linguistic clues from the first part of the sentence, the participant has no other choice than to rely on a general context of his own to come up with a sensible prediction. This correspondence between the global context and the upcoming word would be derived from these same heuristics and expectations about the regularity of language we already discussed.

The fact that the context could contribute to the linguistic prediction should not be surprising because we already have empirical evidence supporting the idea that a person can have expectations about upcoming information beyond the level of the word.

For example, Rohde et al. (2011); Rohde & Horton (2014) showed that the interpreters (or the hearers) were able to build up expectations regarding the coherence relations of upcoming sentences. Similarly, Brothers et al. (2017) measured the effect of the contextual influences coming from the different instructions given to the participants for the same task: one set of instructions emphasized the task's predictive nature while the other did not. Their results provided evidence that predictions do not depend solely on the nature of the stimuli presented to the participant but also on the experiment's contextual setting.

Finally, more empirical evidence for a prediction made above word-level comes from studies on turn-taking that argue that one must anticipate the end of a speaking turn to prepare her response in advance Levinson (2016); Levinson & Torreira (2015); Riest et al. (2015). In such cases, it is the whole sentence that is anticipated and not only the word form.[105]

What transpires from these results and this discussion is that the influence from the context is significant and that it might play a facilitating role in linguistic prediction. Crucially, contextual contributions must not be relegated to the background as they are fundamental for understanding language comprehension and, in our case, linguistic prediction (Hasson et al., 2018). In this chapter, my goal is to describe the nature of these contextual influences, how they could be modeled and combined with the contribution from the meaning compositions we discussed in Chapter 4.

5.2 Representing Context

To model the mutual effects between the context and a sentence, we need to represent the context to compute its contribution during linguistic prediction. This section discusses two kinds of information that contribute to the derivation of this contextual representation: the Situation Model and the Topic Model. These two kinds of representation naturally correspond to the distinction we discussed in

[105]The same kind of effect about predictions derived from higher-level expectations was described by Karuza et al. (2017) when they asked participants to generated expectations about upcoming images that were presented to them. When these images were presented in a structured way that allowed the participant to form a higher-level representation of this network architecture, the uncertainty about the upcoming images was minimized, and the reaction time decreased accordingly.

Chapter 3 about different kinds of coordination (Gärdenfors, 2014): coordination about the meaning of a term, which correspond to the topic model, and the coordination of knowledge about the state of the world, which correspond to the situation model. After discussing the nature of these two models, I present a multi-layered representational approach for modeling linguistic prediction.

5.2.1 Situation Model

The origin of the situation model lies in the approach put forward by Bransford et al. (1972) where the interpretation of a sentence is linked with the derivation of a model about the state of the world, and this situation model is then stored within the long-term memory of the hearer.[106] A situation model is an integrated mental representation of a particular state of the world (Zwaan, 2008). When creating or updating a situation model, we have to consider three kinds of information that can be represented: the situational framework, the situational relations, and the situational content (Zwaan, 2016). The situation framework is related to the spatio-temporal anchoring of the situation, while the situational relations are about the representations of the relationships between the different events expressed by the sentences. The situational content contains information about the entities and their features.

For example, when reading (116), we are deriving a situation model containing two entities and some basic features about *Jonas* being animate and *bottle* being inanimate. We also derive information about the event described in the sentence,

[106]The Situation Model was first introduced by van Dijk & Kintsch (1983) around the same time as the Mental Models of Johnson-Laird (1983).

namely, Finding(Jonas,bottle). We might also have a spatio-temporal time-stamp, i.e., January 21st, 1987, for example.

(116) Jonas finds a bottle.

Situations have to be distinguished from schemata that are type-representations of stereotypical situations. Schemata usually contained the entities present in the situation, i.e., the objects used and the actions that are stereotypically performed at this location, while situations models are token-representation about a specific moment at this location (Zwaan & Radvansky, 1998). To illustrate this difference, if we set the location to be a school. Then schemata representation would be about the different people at school, the different kinds of objects we could use, the different actions that could be performed, whereas the situation model would be about a particular moment at this school, e.g. "I was in school on December 20th with Ayesha in D-3323".

One way to represent these situation models is to use the DRT boxed representation. In Chapter 3, we argued that DRT fulfilled our requirements for a theoretical model of linguistic prediction; it is thus natural to use discourse representation structures (DRSs) to represent situation models. In Table 5.1, we see the representation of a situation for the sentence *Jonas finds a bottle*.

At any given time t_0 during, a situation model contains all the entities and events that were expressed at time $t_{<0}$. The nature of the representation of the situation model is very close to what is usually thought of when defining the context of enunciation, i.e., the time, location, and information about the state of the world and the entities involved in the current situation.

Table 5.1 Representation of a situation model in the DRT boxed format

Jonas finds a bottle.
x : Jonas(x)
y : bottle(y)
$finds(x, y)$

To illustrate the influence the situation model has on our expectations regarding upcoming information, we can go back to the example we presented in Chapter 3. Without any context, is it difficult to predict the next word in (117-b) because the number of possibilities is very large. However, if we add a situation model like (117-a), it becomes much easier to predict that the next word will probably be *chess*.

(117) a. $[x, y, z, : John(x), chess\_board(y), chess\_tournament(z), Being\_at(x, z)]$

b. John plays ...

In this situation, the local context *John plays* is not constraining enough, so we need the contribution from the situation model to limit the number of possible answers. This example shows that a situation model can influence the results of a linguistic prediction. However, some cases could not be explained by using a situation model, and we need a different kind of contextual information, i.e., we need a topic model.

5.2.2 Topic Model

Topic models are based on the idea that a document is an amalgamation of different topics where each topic is itself a probability distribution over different words (Steyvers & Griffiths, 2010). Topic models are statistical (or probabilistic) models used in machine learning and natural language processing to discover the topics occurring in a document or a series of documents (Blei et al., 2003). The term *topic* should not be confused with the concept of *topic* usually used in linguistics when discussing information structure within a sentence (Halliday, 1967). In linguistics, the information structure comprises the basic notion of *focus*, *givenness* and, most importantly, the *topic* which specifies what the statement is about.[107] When it comes to the topic model used here, a topic is simply a probability distribution over the words present in a document, and it is not explicitly linked with the information structure at the level of the individual sentences.

According to the topic model approach, to interpret a sentence, an interpreter has to retrieve the referred concepts from long-term memory (Kintsch, 1988), and this retrieval process is facilitated when the interpreter uses a representation of the context to help her predict related concepts and disambiguate word meanings (Griffiths et al., 2007, p.211). It is possible to represent this contribution from the context using several different methods: associative semantic networks (Collins & Loftus, 1975), similarity spaces (Gastaldi, 2020), or probability distribution between a word and a topic (Steyvers & Griffiths, 2010). These three kinds of representations are illustrated in Figure 5.2 (Griffiths et al., 2007, from their Figure 1).

[107] See Krifka (2008) for a thorough introduction of these three notions.

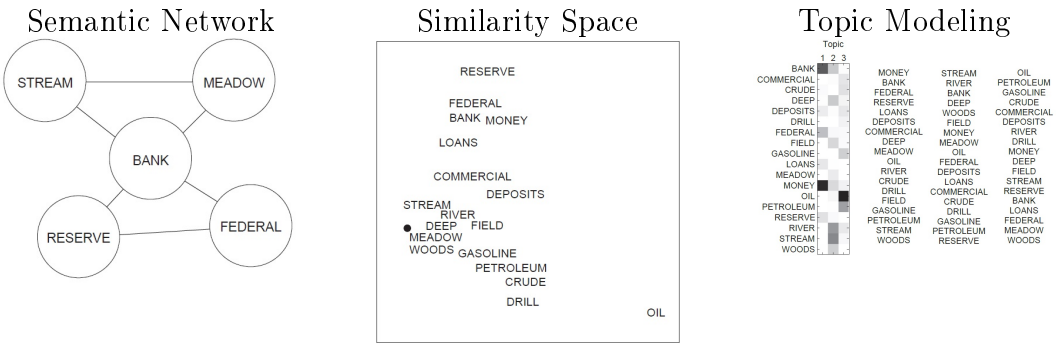


Figure 5.2 Three ways to represent semantic information

Both semantic networks and similarity spaces have difficulties tackling topic modeling because they both involve unipartite graphs. A unipartite graph is a graph where all nodes are of only one kind and which can all be connected indiscriminately, as it is illustrated on the left side of Figure 5.3. This limitation to a unipartite graph is problematic because it means we cannot distinguish between the nodes. In Chapter 4, we derived particular similarity spaces at the RB-level and the P-level because their respective linguistic units were not part of the lexical-level similarity space. To capture the contribution from the context, we also need to have a graph with two kinds of nodes to distinguish the level of the words from the topic-level (Griffiths et al., 2007).

A bipartite graph consists of two disjoints sets of nodes, and the only connections that are allowed are between these different kinds of nodes (Rosen, 2012). A bipartite graph is illustrated on the right side of Figure 5.3. Topic modeling needs this hierarchically structured graph to represent the difference in representational level between the words from a sentence and the topic to which they are linked.

The representation of a topic is a probability distribution over different words (Blei et al., 2010). Conversely, each word is probabilistically linked with a series

Figure 5.3 Left-side: Unipartite graph. Right-side: Bipartite graph

of topics via a latent structure (Steyvers & Griffiths, 2010). To differentiate between this latent structure and the level of the words, we need a bipartite graph representation as in 5.4 (Griffiths et al., 2007, adapted from their Figure 4).

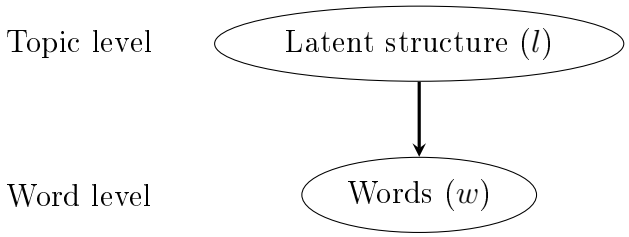


Figure 5.4 Latent structure and Generating model at word-level

The latent structure l, i.e., the topic model, generates words w using the conditional probability distribution $P(w|l)$. This way, a word is more likely to be generated if it is more probable within that topic and vice-versa. We can illustrate this with an example given in Griffiths et al. (2007). Suppose we have a high probability for generating the words *woods* and *stream*. In that case, the topic is likely to be 'countryside.' Similarly, if we have a high probability of generating the words *federal* and *reserve*, then the topic is probably 'finance.' Conversely, if we know that the topic represented in the latent structure is 'finance,' then we

would have a higher expectation regarding the production of a word like *bank* over a word like *woods*. We could summarize all this using the probability that a word w_i is produced within a document, given a topic z and several topics T (Steyvers & Griffiths, 2010).

(118) $P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j_i)P(z_i = j)$

In (118), the term $P(z_i = j)$ refers to the probability that the jth topic is sampled from the ith word and $P(w_i|z_i = j_i)$ is the probability of word w_i under topic j.

Polysemy One great feature of these topic models is their ability to disambiguate lexical meaning (Griffiths et al., 2007) or, put differently, to capture polysemy (Steyvers & Griffiths, 2010). This disambiguation is made possible because it is possible to infer the meaning m of a word w given a topic z (Griffiths et al., 2007).

(119) $P(m|w, z) = \frac{P(w,m|z)}{\sum_m P(w,m|z)}$

Following the topic model approach, the polysemy of a word is represented as uncertainty over possible topics (Griffiths et al., 2007), and it is this uncertainty over topics that allows solving the ambiguity about different meanings of a word. For example, the polysemy of the word *play* is depicted in Figure 5.5 with respect to three different topics: *play* music, theater *play*, *play* games.[108]. In Figure 5.5 (Steyvers & Griffiths, 2010, Figure 9), we see that the three senses of *play* have a relatively high probability for all their respective topics.

[108]This figure was obtained from a 300 topic solution for the TASA corpus, and it is taken from Steyvers & Griffiths (2010)

Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

Figure 5.5 Three topics related to the word *play*

5.2.3 Multi-layered Representation of Linguistic Prediction

To model the contribution coming from these two models, we have to understand the different levels of representations involved in linguistic prediction (Berkum, 2013) and the relationship between them. The first thing to do is treat the sentence level separately from the context level using a bipartite graph. From there, we can then build a multi-layered representational view of linguistic prediction as illustrated in Figure 5.6.

Even though incrementality is not taken into account by the model at this point, it is still important to note that the temporal relationship between these two different levels is crucial as the representations for the situation and the topic models at time t are based on the linguistic information processed at time t_{i-1}. The sentence-level representations are crucial contributors to the derivation of these contextual-level representations. In turn, the contextual representations constraint the way these sentence-level representations are interpreted at time

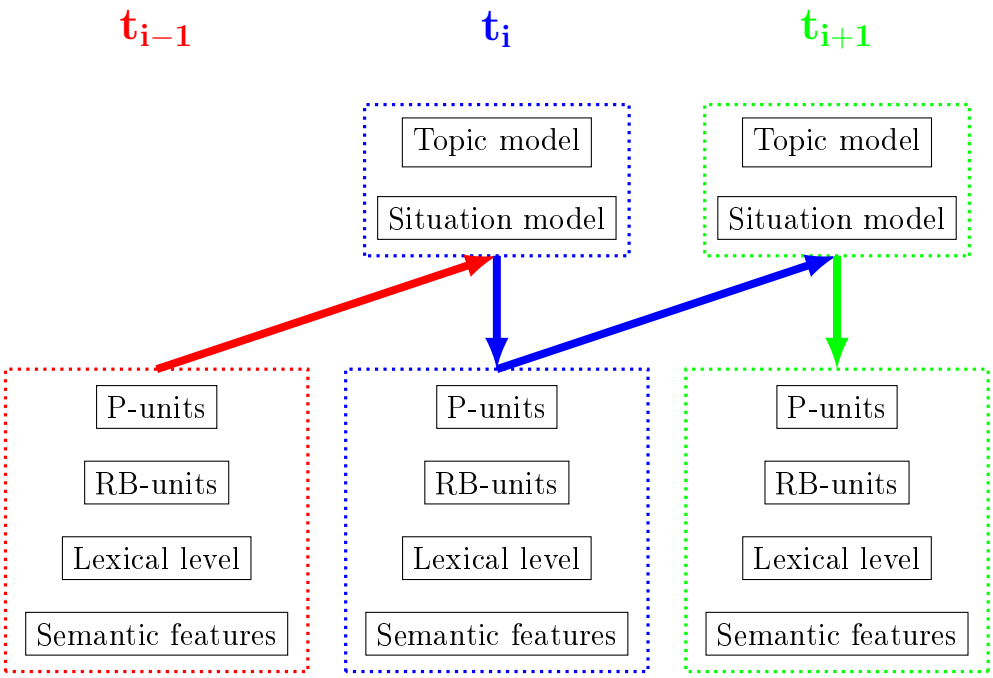


Figure 5.6 Temporal Depiction of the Co-dependence between the sentence-level representations and the contextual representations

t_{i+1}. In other words, the contextual representations and the interpretation of perceptual inputs are constraining each other (Zwaan, 2016). This compositional-contextual processing loop is depicted in Figure 5.6. This thesis assumes that the mapping between the contextual-level representations and the sentence-level representations is mediated through coordination.[109]

Figure 5.7 integrates two contextual representations we just described and the four sentence-level representations we discussed in Chapter 4. In this thesis, I distinguish between two kinds of mapping functions: the functions that derive the

[109]A hierarchical view of language processing is consistent with many results from visual scene recognition studies where the participants use global features rather than local features to describe a scene (Oliva & Torralba, 2001; Oliva, 2005). In such scenes, the global features of the scene, i.e., the gist of the scene, influences the way local predictions are made (Torralba & Sinha, 2001).

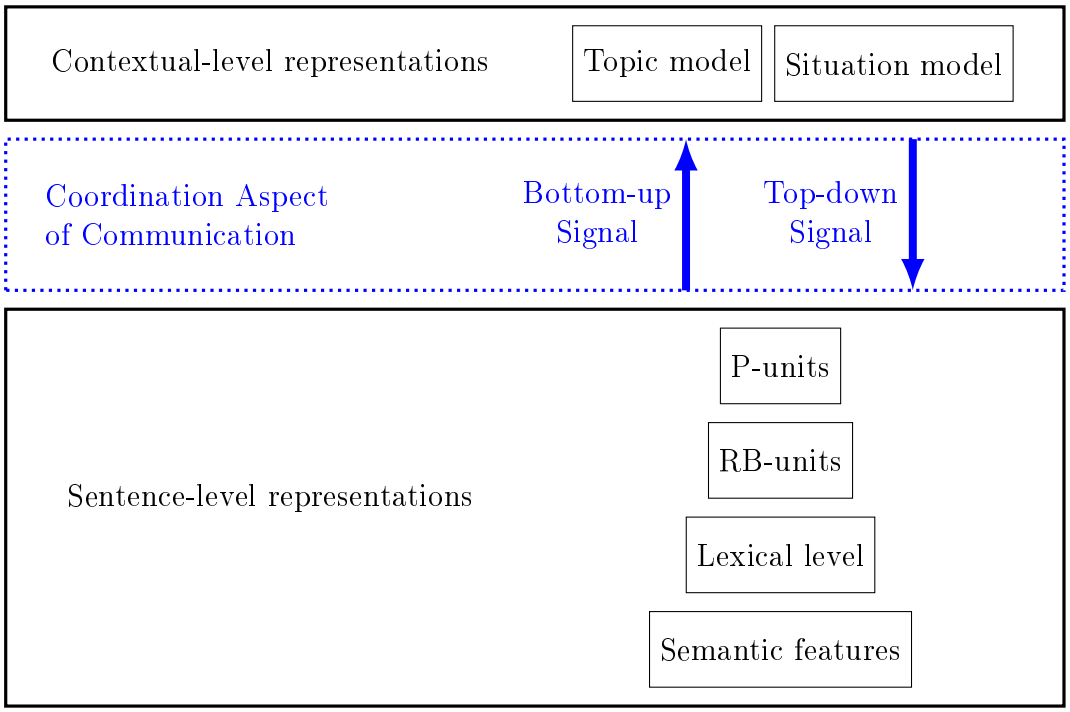


Figure 5.7 A Multi-layered Representations of meaning

contextual level from the sentence-level representations and the mapping function corresponding to the influence of the contextual level representations imposed over the sentence-level linguistic prediction. These mapping functions are respectively associated with a bottom-up signal and a top-down signal, and these two signals represent the contribution from the coordination aspect of linguistic communication. The critical thing to retrieve from Figure 5.6 is that both mapping functions play a role when deriving a linguistic prediction.

The bottom-up signal is responsible for the constant update of the contextual level. This signal is called *bottom-up* because it goes from the sentence level up to the contextual level. The top-down signal represents the constraints imposed on the sentence level by the expectations derived at the contextual level. It is called a top-down signal because it starts at the top and goes down in the representational

hierarchy. These two signals involve different kinds of coordination: the bottom-up signal is about the coordination between two agents, whereas the top-down signal is about the coordination between the contextual representation and the sentence-level representation.

It should also be mentioned that intra-representational level influences have been deliberately omitted from Figure 5.7. This means the possible interactions between the four different representations at the sentence level and those between the two contextual representations are not taken into account in this approach. It would be interesting to investigate these intra-representational level influences, especially when considering an incremental linguistic prediction model. However, it is outside the scope of this thesis.[110]

Finally, the terms *bottom-up* and *top-down* are often used in cognitive science and linguistics, but it is rarely defined what these refer to (Rauss & Pourtois, 2013). Before presenting how to model these bottom-up and top-down signals, it is crucial to specify how these signals are interpreted with respect to this multi-layered representation. These two notions are often centered around the perspective that bottom-up processing is transforming a lower-level representation into a higher-level representation, and top-down processing is doing the contrary (Palmer, 1999). To have bottom-up and top-down signals, we need to process information hierarchically at different representational levels, and the exchange of information between these levels need to be bidirectional (Rauss & Pourtois, 2013). However, different hierarchies can have different structural and functional flexibility. Engel et al. (2001) distinguished between four families of hierarchies: anatomical, cognitivist, gestaltist, and dynamicist. In the anatomical hierarchy,

[110]Notwithstanding the direction of information flow, these influences would not strictly qualify as top-down or bottom-up signals because they do not involve a coordinated representational transition as the one between the sentence-level and the contextual level representations.

the top-down and bottom-up processes are strictly functional, which means that the different levels are linked with a processing connection. In the cognitivist hierarchy, bottom-up and top-down signals respectively correspond to stimulus-driven and expectation-driven processing. According to the gestaltist view, top-down processes are contextual modulations of bottom-up processing. Finally, in the dynamicist hierarchy, large-scale dynamics can have a predominant influence on the local processing of information.

In the multi-layered representation presented here, the top-down and the bottom-up signals are of the dynamicist type, and it is compatible with the idea that the information is processed simultaneously at all representational levels.[111]

5.2.3.1 Top-down Signal

Top-down influences are influencing the lower-level representations by modifying the expectations of the hearer. For example, if we go back to the truncated sentence *John plays...* and we assume that it is uttered when the two interlocutors are at a chess tournament and discuss chess players in the room.

(120) John plays...

a. chess.

b. cello.

The top-down signal acts as a constraining signal that influences the probability

[111] I come back to this issue in Chapter 6 when discussing predictive processing architectures.

distribution of potential continuations. For example, if we know we are at a chess tournament via the situation model, then the continuation (120-a) would be highly probable. However, if we also know that at that particular moment, the topic of our discussion was the musicians we knew, then the combination of the topic model and the situation model would transform (120-b) into the most probable continuation. To model these kinds of contextual constraints, I propose a probabilistic approach that can encompass both the effects from the topic model and the ones from the situation model while being compatible with the model I presented in Chapter 4.[112]

5.2.3.2 Bottom-up Signal

Now that we have discussed the top-down influences from the contextual representational level, we must also acknowledge how this level is derived in the first place. As we saw in Figure 5.6, the content of the situation model and the topic model at time t_i is constructed from the linguistic content expressed by the sentence-level representations at time t_{i-1}. Put differently, when we are presented with linguistic information in the form of a truncated sentence at time t_{i-1}, this information is interpreted and is used to derive a topic model and a situation model at time t_i.

At the beginning of this chapter, we mentioned that a hearer could usually understand a speaker even for cases when the interpretation of the sentence is inherently uncertain because of the disambiguating effect coming from the context. For example, in Chapter 3, we presented a case of referential uncertainty between a *blue*

[112]To model the top-down influences, it is also possible to use a topological approach based on the geometry of the conceptual space. This other approach is discussed in Chapter 6.

square and a *green square* where the hearer had to rely on the mapping function of the speaker to retrieve the correct reference.

Similarly, to use the topic model, we have to take into account that these models are derived from the sentence-level representations, which are themselves produced by the speaker. This implies that we must also retrieve the speaker's intended meaning if we want to derive the correct topic model.

This problem about the ambiguity of meaning, referential or otherwise, is not related to the situation model or the model per se, but it must be taken into account at the lexical level when deriving a particular contextual model. One way to do this is to use a built-in representation of the speaker's perspective to retrieve the intended meaning for every linguistic unit. Integrating this speaker's perspective during the interpretation process could be performed using the same kind of game-theoretic process described in Chapter 3. In this chapter, I use probabilistic approaches based on game theory to model this bottom-up signal.

5.3 Modeling Coordination

In the last section, I have presented a multi-layered view of the representations involved in the derivation of a linguistic prediction. I mentioned that the bottom-up signal and the top-down influence between the contextual and compositional levels were modeled probabilistically. This section introduces the Bayesian framework and discusses its importance for the modelization of different linguistic phenomena. I then present how to use a Bayesian approach to model both bottom-up and top-down signals.

5.3.1 Bayesian Framework

The Bayesian Framework's core idea is that inferences are probabilistic (Jones & Love, 2011), which implies that given a set of premises, a consequence is drawn by using probabilistic knowledge about the relationship between these premises and this conclusion. Bayesian inferences are to be distinguished from other kinds of statistical inferences like frequentist inferences. According to the frequentist approach, an inference is thought to be related to the number of occurrences and the potential repeatability of an event, while the Bayesian inference is about modeling the uncertainty of an outcome (Chater et al., 2006b). When considering a series of experiments involving repetitive manipulation, such as a series of coin tosses, one can use a frequentist inference to guess the subsequent outcome. In the case of Bayesian inferences, the inference is derived from a subjective conception of probability, meaning that the inference is about degrees of beliefs about the possible outcomes, and it is not be based on the frequency of this event (Chater et al., 2006b; Jones & Love, 2011). For example, two persons might have different subjective probabilities for the same event because their prior knowledge about such events is different (Chater et al., 2006a). Bayesian inference is thus a rational way to account for uncertainty about the state of the world around us (Oaksford et al., 2009).

Treating Bayesian inferences as inferences about uncertainty has some significant advantages because complex probabilistic models can now be worked out using powerful computational tools developed in such fields as statistics, machine learning, and artificial intelligence (Griffiths et al., 2008, p.3). It offers a potentially unifying framework to help better understand many different aspects of cognition by modeling them as inferences about uncertainty (Jones & Love, 2011). In re-

cent years, Bayesian approaches have been used to address a plethora of different problems coming from different disciplines in cognitive sciences: inductive learning and generalization (Tenenbaum et al., 2006), language acquisition (Chater & Manning, 2006; Xu & Tenenbaum, 2007), symbolic reasoning (Oaksford & Chater, 2001), motor control (Körding & Wolpert, 2006).[113] In each of these fields, the probabilistic nature of these Bayesian inferences about the uncertainty of an outcome remains the same, but the structures that are being inferred are specific to the field in question (Chater & Manning, 2006).

Modeling empirical data using a Bayesian approach is not necessarily correlated with the idea that the human brain is itself a Bayesian machine. Most people have some difficulties with textbook probability problems that are not related to assessments of causal efficacy or causal models (Chater et al., 2006a; Oaksford et al., 2009), and, even though this idea is still disputed (Chater et al., 2006a; Jones & Love, 2011), it is often argued that people are not purely Bayesian, but only approximately Bayesian (Jacobs & Kruschke, 2011).

It has been argued that Bayesian approaches might not be the most natural way to offer a mechanistic explanation of how a biological system works (Martin & Doumas, 2017; Griffiths et al., 2008), but arguing that people are approximately Bayesian should not be construed as a strong constraint at the implementational level of analysis as defined by Marr (1982). Furthermore, even at the algorithmic level, there is no firm commitment as to how these probabilities are computed in the first place (Oaksford et al., 2009). Besides, most Bayesian approaches are meant to be formulated at the computational level (Griffiths et al., 2008), and they are simply interpreted as a rational way to model uncertain inferences.

---

[113] A probabilistic approach has also been developed to explain empirical results for the famous Wason's selection task (Oaksford & Chater, 1994, 1996).

5.3.1.1 Bayesian Model

A prediction is nothing more than the expression of a belief about a state of the world (Norris et al., 2016), and Bayesian inferences are used to predict this state of the world given a particular global context. To predict the state of the world given a particular set of conditions, we need to use our beliefs about the current state of the world and our beliefs about the relationship between this state and other potential states of the world. In other words, we need to know the prior probability and the likelihood function (Trapp et al., 2016).

To better distinguish these two notions, we use the concept of hypothesis space, where every hypothesis acts like potential states of the world. Here, the term *hypothesis* refers to a probability distribution, and it is not related to the concept usually used in psychology, which has nothing to do with explicit reasoning (Jones & Love, 2011). The likelihood is the probability that a particular observation or feature is present in the world, given a particular hypothesis, and it can be written as in (121).

(121) $P(\text{observation}|\text{hypothesis}) = P(o|h)$

The other important component of the Bayesian model is the prior probability distribution representing one's beliefs about each hypothesis. This prior probability distribution about a specific hypothesis can be written as $P(h)$, and it should be thought of as being independent of any observation, i.e., what is my belief regarding the state of the world without any new sensory input (Griffiths et al., 2008). For example, to infer whether a new input is consistent with what we expect, e.g., "a white ladder all covered with water", we have to consider both the

prior probability of a ladder being covered with water, but also the likelihood that such as situation is happening given the actual state of the world (Trapp et al., 2016).

Together, the prior and the likelihood are the two main components of the Bayesian Model (Jones & Love, 2011), and we can combine them using Bayes' Theorem or Bayes' rule (Bayes, 1764) to obtain the posterior probability as in (122). Crucially, Bayes' rule itself does not specify how the priors and the likelihood are obtained, so it does not forbid us to use symbolic techniques to estimate these probability distributions (Zeevat, 2015b).

(122) $P(h|o) = \frac{P(o|h)P(h)}{P(o)}$

According to (122), the posterior probability of having a belief that the world is in the state h given a particular sensory input or observation o is proportional to the prior probability of being in a world h and the likelihood of encountering an observation o within this world h. The term $P(o)$ is the prior probability related to the observation o, and it could be rewritten as the sum of the other variables in a joint distribution: $P(a) = \sigma_b P(a, b)$. This last operation is called *marginalization* and is made possible by using the fact that a marginal probability of a property a is equivalent to the probability distribution when the other variables are not taken into account. Using this transformation, we can rewrite Bayes' rule as in (123) (Griffiths et al., 2008; Tenenbaum et al., 2006), where $\mathcal{H}$ is the set of all hypotheses about the state of the world. The denominator in (123) acts as a normalization factor.[114]

[114]The normalization factor is there to make sure that the sum over all possible outcome is equal to 1

$$P(h|o) = \frac{P(o|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')} \tag{123}$$

In summary, a Bayesian model is to be viewed as a method to derive a posterior probability distribution from the prior probability distribution and the likelihood probability distribution. To illustrate this process, we can look at the Figure 5.8 (Körding & Wolpert, 2006, Fig.I) where x acts as our hypothesis h and we see that combining our prior and our likelihood gives rise to a probability distribution that has weighted the hypotheses by their likelihood (Jones & Love, 2011).

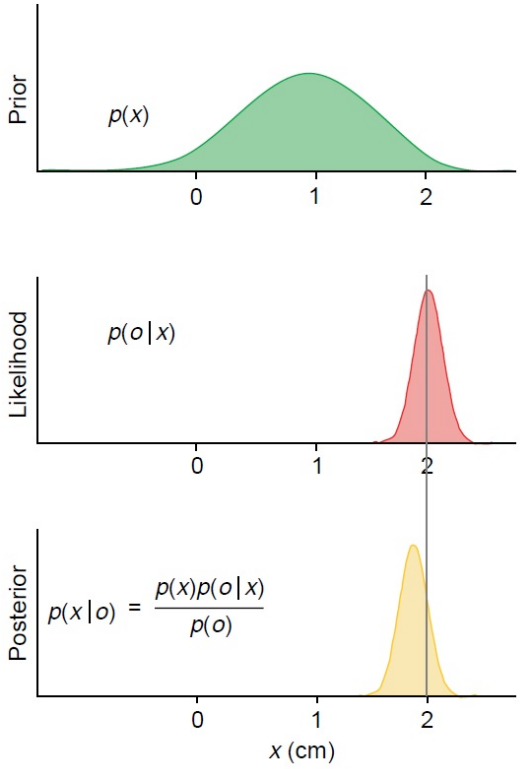


Figure 5.8 A schematic depiction of the Bayes's rule

5.3.1.2 Rational Speech Act (RSA) Model

An exciting domain of application for the Bayesian framework is pragmatics. Chapter 3 discussed different approaches to pragmatics, namely Gricean pragmatics, Relevance Theory, and Game-theoretic pragmatics. Here I present the Rational Speech Act (RSA) model, which uses a Bayesian approach to model linguistic coordination between two agents during communication (Goodman & Stuhlmüller, 2013; Goodman & Frank, 2016; Bergen et al., 2012; Frank & Goodman, 2012; Franke & Jäger, 2016; Scontras et al., 2018).[115] Similarly to these other approaches in pragmatics, Bayesian pragmatics conceives language interpretation as "rational inference based on an intuitive theory of language production" (Goodman & Stuhlmüller, 2013, p.174), and it sits close to Game-Theoretic pragmatics because it is interested in modeling behaviors in terms of reasons and purpose without the recourse to communicative maxims (Franke & Jäger, 2013). Instead of basing their assumptions on maxims, Bayesian approaches of pragmatics assume that the listener considers that the utterance is chosen optimally by the speaker. The listener then interprets this utterance probabilistically using a Bayesian inference (Goodman & Stuhlmüller, 2013).

To better understand the relationship between Game-Theoretic pragmatics and Bayesian Pragmatics, we can recall the signaling game we discussed in Chapter 3. In a signaling game, we have a speaker, a listener, and a conversational move that consists of an utterance produced by the speaker. A signaling game could thus be thought of as a context of enunciation, and the goal of the game is for the listener

---

[115] What I describe here as Bayesian pragmatics is often categorized as probabilistic pragmatics (Tessler & Goodman, 2014; Franke & Degen, 2015; Franke & Jäger, 2016) because it involves the use of probabilities.

to retrieve the new state of the world from her prior knowledge and this utterance. In Game-theoretic pragmatics, the pragmatic reasoning required to retrieve the state of the world is modeled as game-theoretic solution concepts (Franke, 2013). In Bayesian pragmatics, these inferences are modeled using probabilistic tools and Bayesian statistics (Frank & Goodman, 2012), and this formal framework is called the Rational Speech Act theory of language understanding or RSA model (Goodman & Stuhlmüller, 2013; Goodman & Frank, 2016).

RSA model's central tenet is that a speaker chooses to produce an utterance parsimoniously while taking into account other alternatives in a given context (Goodman & Frank, 2016). This optimality of the speaker allows the listener to retrieve the interpretation of the sentence by inverting the model of the speaker (Goodman & Stuhlmüller, 2013). Under this view, to interpret an utterance, we have to go through a recursive social reasoning (Bergen et al., 2012) about the speaker's perspective (Bergen et al., 2012), much like what was described for Game-Theoretic pragmatics in Chapter 3. Once this recursive social reasoning is accomplished, the listener can use the speaker's beliefs to derive the interpretation of the utterance and then update her own beliefs accordingly via a Bayesian inference (Goodman & Frank, 2016).

This iterative and recursive inversion of perspective requires a back-and-forth reasoning about each other's point of view and each other's beliefs about the state of the world (Franke, 2009). In RSA terms, the listener uses Bayesian inference to retrieve the speaker's intended meaning given the utterance he produced (Frank & Goodman, 2012). This recursive approach is sometimes called intentions-first approaches to pragmatics, or iterated X-response (IxR) models (Franke & Jäger, 2013). These IxR models were first developed to capture epistemic effects, such as taking into account the belief of the speaker (Goodman et al., 2009) but they are close with the RSA model. The main difference between the two comes from the

fact that the IxR models are primarily interested in modeling actions performed by the listener after processing an utterance. In contrast, RSA is instead focussed on modeling the speaker's beliefs that help the listener retrieve the interpretation of this utterance (Franke & Jäger, 2013).

To understand how this recursive social reasoning works, we first consider the literal (or naive) listener's perspective L_0. This literal listener uses Bayes' rule to update her beliefs while assuming that the literal meaning of the utterance u is true (Goodman & Frank, 2016). In (124), we have that the conditional probability that the state of the world s is given the utterance u is proportional to the product of the semantic denotation of u evaluated at s which is written $[\![u]\!](s)$ and the probability of being in the state s in the first place (Scontras et al., 2018).

(124) $\quad P_{L_0}(s|u) \propto [\![u]\!](s) \cdot P(s)$

When a speaker chooses to utter a particular utterance, he might produce the easiest utterance without any concern to the person that has to understand it, but, generally speaking, the speaker has to take into account that the literal listener will try to retrieve the meaning of this utterance using (124). This more sophisticated speaker is called a pragmatic speaker because he considers the listener when choosing his utterance. Another way to put this is to say that the pragmatic speaker wants to minimize the effort that the literal listener must exert when retrieving the world's state from the utterance (Scontras et al., 2018). Mathematically, the conditional probability that a pragmatic speaker chooses one utterance in particular from the set of all alternatives is expressed as being proportional to the exponential of the Expected Utility function U_{S_1} as we see in (125) (Goodman & Stuhlmüller, 2013).

(125) $P_{S_1}(u|s) \propto exp(\alpha/U_{S_1}(u;s))$

In (125), the α is a parameter that captures the rationality of the speaker, and the Expected Utility function U_{S_1} is to be thought of as the weighted average of all the utilities, i.e., the desirability of a given option, that a speaker expects when choosing one alternative over another (Franke & Jäger, 2016; Bergen et al., 2012). This Expected Utility function is given in (126), where $C(u)$ represents the cost of producing the utterance u.[116]

(126) $U_{S_1}(u;s) \propto logP_{L_0}(s|u) - C(u)$

The Expected Utility is thus proportional to the logarithmic probability that the literal listener retrieves the state of the world s from the utterance u minus the production cost. Now that we have presented the pragmatic speaker, we must also consider that real-world listeners are usually not literal in that they also consider the speaker's perspective when interpreting utterances. In other words, we need to have a pragmatic listener that will use Bayes' rule to infer the probability of a state of the world s from the utterance u from the speaker's perspective.

(127) $P_{L_1}(s|u) \propto P_{S_1}(u|s) \cdot P(s)$

In (127) we have that this probability is proportional to the probability that the pragmatic speaker chooses u given s times the prior probability that the world is itself in the state s.[117] Putting all these together, we get that the RSA model

---

[116] At this point, I am not taking into account the cost $C(u)$, but I come back to discussing processing costs in Chapter 6.

[117] Please see Franke & Jäger (2016); Goodman & Frank (2016); Scontras et al. (2018) for

framework offers a Bayesian recursive structure to understand the interpretation process during a coordinated interaction between two agents.

When a speaker wants to utter something to convey a given state of the world, he chooses an utterance that will be easy to interpret for a literal listener. When the chosen utterance has been produced, the listener takes this perspective into account by interpreting the utterance chosen by the speaker within a set of possible alternatives.[118] To better illustrate the difference between the literal listener and the pragmatic listener in the RSA model, we can consider a Reference Game where the goal is to retrieve the reference that the speaker intended. In this example, the speaker can only produce one word, and only three objects can be referred to, as in Figure 5.9 (Frank & Goodman, 2012, adapted from their Fig.A).

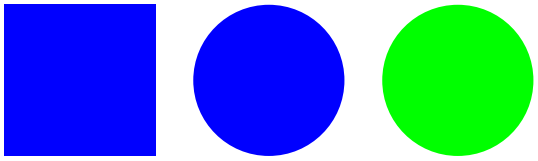


Figure 5.9 Three objects that can be referred to

Let us first consider that the word that the speaker had uttered is *green*. Then it will be easy for the listener to retrieve the correct referent intended by the speaker because there is only one possible referent that is green. On the other hand, if the word that had been uttered is *blue*, the literal listener interpretation would be based on the truth value about the possible referents and, because there are two

---

a more detailed explanation of the role of the utility function and all the motivations behind the pragmatic speaker and the pragmatic listener as well as a presentation of the normalization constants needed to transform these proportional relations into equalities

[118] In the RSA model, utterances are usually taken into account as a whole. More recent implementations tackle incremental processing of utterances (Cohn-Gordon et al., 2019), but those approaches are outside this thesis's scope.

possible referents, the literal listener attributes a 50/50 probability to both these options, namely the blue circle and the blue square.

If we keep the same word *blue*, but we instead consider the pragmatic listener, the outcome is different. According to the pragmatic listener, the speaker chose to utter *blue* instead of other possible words to refer to a particular object. The word *blue* is compatible with two objects, but when considering the perspective of the speaker, the pragmatic listener infers that if the speaker intended to refer to the blue square, him uttering *square* would be a more rational choice because it disambiguates the interpretation. Therefore, that the speaker has chosen to utter *blue* probably means that the referred object is the blue circle. Depending on the parameter α that we presented earlier, the pragmatic listener would attribute a higher probability to the blue circle than the blue square. If we increase the value of α, i.e., the speaker is becoming more optimal, the probability ratio $P(\text{blue circle})/P(\text{blue square})$ will tend to infinity because the probability of referring to the blue square will decrease drastically. On the other hand, if we decrease the value of α, i.e., the speaker is becoming less optimal, the probability ratio $P(\text{blue circle})/P(\text{blue rectangle})$ will revert to 1 as was the case for the literal listener.

Here we have only presented a simple example of a Reference Game, but the RSA model does not simply hypothesize about simple linguistic interpretation cases. It gives us a whole spectrum of variations that can be explored by modifying and adding more parameters while maintaining the same Bayesian structure (Goodman & Frank, 2016). Many recent implementations of the RSA have tackled such issues like Scalar Implicatures (Goodman & Stuhlmüller, 2013; Franke & Bergen, 2020), Free-Choice Inferences (Champollion et al., 2019), and non-literal and figurative language by parametrizing the model to include uncertainty at different levels, but these implementations are beyond the scope of this thesis.

5.3.2 Modeling Bottom-up Coordination

In our multi-layered representation depicted in Figure 5.7, we had that the bottom-up signal was the signal that derived the contextual representations, i.e., the topic model and the situation model from the sentence-level representations.

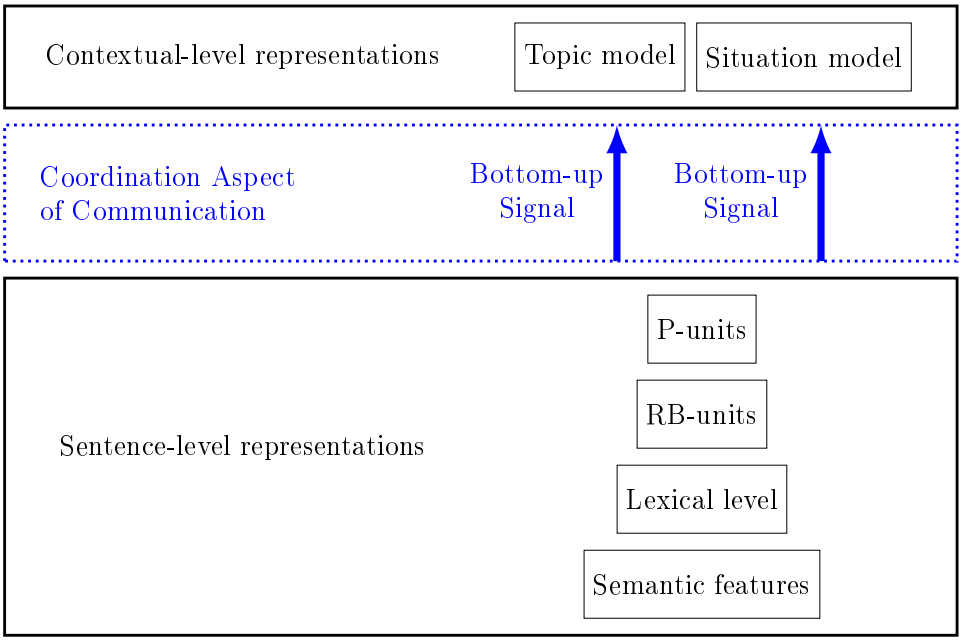


Figure 5.10 A depiction of the Bottom-up signals

As shown in Figure 5.10, we have to consider two bottom-up signals, one for each kind of contextual representation, and I treat each topic model and situation model bottom-up signals separately.

5.3.2.1 Situation Model

In the last section, we described the situation model as a model of the entities and the events expressed by a sentence. The derivation of such a model requires recognizing and storing information about these entities and events. To do so, we can use a DRT-like implementation to derive a Discourse Representation Structure (DRS) and use it as our situation model. For example, following the DRT format given in Bos (2015), the sentence in (128) would be represented by the DRS given in Figure 5.11.

(128) John grabs the donkey.

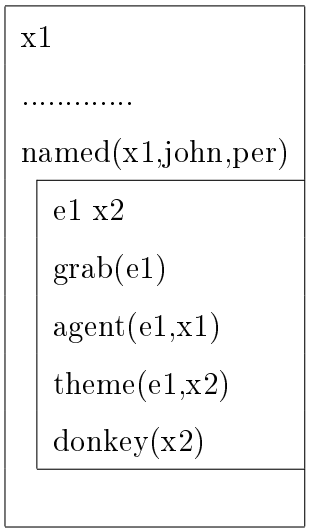


Figure 5.11 DRS for the sentence "John grabs the donkey."

In Figure 5.11 we can see there is a person x1 named "John" and there is an event $e1$ involving an entity $x1$ (John) and an entity $x2$ which is a donkey. This DRS gives us the thematic role labeling for the entities involved in this event, $e1$, which,

in this case, is that event $e1$, $x1$ is an agent, and $x2$ is a theme. If we were to summarize the information presented in this DRS, we would have a list of entities, a list of events, and a thematic role attribution linking the events and the entities. As a whole, this DRS represents the state of the world the listener is in when retrieving the meaning of the following information. In other words, this DRS is tantamount to the contextual level situation model derived by the listener.

To integrate this DRS content into our linguistic prediction model, we have to represent it using word embeddings, much like what we did in Chapter 4. One way to do this is to note that the way this DRS content is structured makes it possible to find correspondences with the four compositional units we previously described: semantic features, lexical level, RB-units, and P-units. The P-units, representing a complete proposition, would correspond to the DRS event, the lexical-level units would correspond to the entities present in the DRS, and the thematic roles could be transposed as RB-units. In the case of the RB-units, every thematic role would correspond to one RB-unit.[119]

In Chapter 4, we presented an example about a cup being taller than a ball, but it would not be too far-fetched to transpose this idea for predicates with thematic roles such as agent and theme. For example, in the case of *grab(John,donkey)*, the RB-units consist of one unit being *grabber+John* and another being *grabbed+donkey*. In addition, the semantic features level would correspond to the property of the entities. In the case that concerns us, *John* is the name of a person, i.e., *person(John)*. With these correspondences in mind, we can rewrite the DRS content

[119]The explanation of the transition from the DRS to the Situation Model is merely to illustrate how this transposition could be achieved, but it is by no means compulsory to start from the DRS described in Figure 5.11 to obtain the Situation Model in Figure 5.12. In this thesis, I choose to represent the Situation Model using the 4 different levels described in the preceding chapter and the focus in this thesis should be on the Situation Model itself rather than on the way it could be obtained.

in terms of different word embeddings.

The situation model described in Figure 5.12 is thus constructed from many distributed representations, and the final step would be to combine them to form only one integrated distributed representation for the whole situation model. Having the situation model represented by one integrated vector is desirable because it makes it easier to model the top-down influence of the situation model on the linguistic prediction.

Entities	$\overrightarrow{\text{John}}$ $\overrightarrow{\text{donkey}}$
Features	$\overrightarrow{\text{person+john}}$
Thematic Roles	$\overrightarrow{\text{grabber+0.5john}}$ $\overrightarrow{\text{grabbed+0.5donkey}}$
Event	$\overrightarrow{\text{grab}}$

Figure 5.12 Detailed Situation Model for the sentence "John grabs the donkey."

The process by which the vector representing the situation model as a whole is derived is reminiscent of what we did when we presented the four different similarity spaces in Chapter 4. Every compositional unit had its own similarity space attached to it, and it allowed us to perform different similarity measures for each of these units. However, it is important to note that all the content and the structure of the Detailed Situation Model as in Figure 5.12 is different from just the sum of all the activation of its parts. For example, if we were to compute the activation caused by all the constituents from the sentence "John grabs the donkey", we would need to take into account the distributed representations expressed in Figure 5.13 in terms of the different compositional units.

P-unit	$\overrightarrow{\text{grabber+john}} + \overrightarrow{\text{grabbed+donkey}}$
RB-units	$\overrightarrow{\text{graber}} + 0.5\overrightarrow{\text{john}}$ $\overrightarrow{\text{grabbed}} + 0.5\overrightarrow{\text{donkey}}$
Lexical-units	$\overrightarrow{\text{grabs}}$ $\overrightarrow{\text{donkey}}$ $\overrightarrow{\text{john}}$
Semantic Features	$\overrightarrow{\text{person}}$ $\overrightarrow{\text{animal}}$

Figure 5.13 Compositional units and their associated distributed representations for the sentence "John grabs the donkey."

Here we kept the semantic features limited to those that are generally relevant in DRSs like features of entities, i.e., *John* is a person and *donkey* is an animal. In Chapter 4, each representational level was computed independently, and they each give rise to a different similarity space, whereas in the case of Figure 5.13, the situation model is taken into account as a whole.

One way to avoid any problem when trying to derive situation vectors is to concatenate the contributions instead of simply adding them together. In Figure 5.13, we see there are four components or sections of the situation model, which means that there will be three concatenations. This way, even if the vectors are added together within each section, these sections will keep their dimensionality. The whole process is illustrated in (129), where $\frown$ represents the concatenation operation.

(129) $\overrightarrow{\text{Situation}} = \overrightarrow{\text{Entities}} \frown \overrightarrow{\text{Features}} \frown \overrightarrow{\text{Roles}} \frown \overrightarrow{\text{Event}}$

In the case of our example about John grabbing the donkey, the word embeddings for the four sections of the DRS can be written as in (130). and the representation of the situation model is the concatenation of these four sections. Thus, the situation vector has the total dimension of these four other vectors, which is $4 \times N$, where N is the dimension of word embeddings.

(130) a. $\overrightarrow{\text{Entities}} = \overrightarrow{\text{john}} + \overrightarrow{\text{donkey}}$

b. $\overrightarrow{\text{Features}} = \overrightarrow{\text{person}} + \overrightarrow{\text{animal}}$

c. $\overrightarrow{\text{Roles}} = \overrightarrow{\text{grabber}} + 0.5\overrightarrow{\text{john}} + \overrightarrow{\text{grabbed}} + 0.5\overrightarrow{\text{donkey}}$

d. $\overrightarrow{\text{Event}} = \overrightarrow{\text{grabs}}$

(131) a. $\dim(\overrightarrow{\text{Situation}}) = \dim(\overrightarrow{\text{Entities}}) + \dim(\overrightarrow{\text{Features}}) + \dim(\overrightarrow{\text{Roles}}) + \dim(\overrightarrow{\text{Event}})$

b. $\dim(\overrightarrow{\text{Situation}}) = (N) + (N) + (N) + (N) = (4 \times N)$

To remain coherent with our linguistic prediction approach from Chapter 4, we used word embeddings to derive the situation model, but it would also be possible to derive it using another approach. For example, Frank et al. (2009) and Venhuizen et al. (2019) used a pre-defined microworld to train a Distributed Situation Space (DSS), where each observation is linked to a proposition. Even if the derivation method differs, these approaches entertain the same kind of ideas that we argued for here regarding the nature and the importance of the situation model in linguistic processing and especially in linguistic prediction.[120]

[120] Venhuizen et al. (2019) also argued for the importance to take into account the situation model in computational models of linguistic processing.

5.3.2.2 Topic Model

As was the case for the situation model, every new linguistic input contributes to the topic model's derivation. When building this kind of topic model, we can use the fact that a topic is itself a probability distribution over different words (Steyvers & Griffiths, 2010). As we described previously, a topic can be represented as a vector where each value corresponds to the probability of encountering a specific word when discussing this topic. For example, in Table 5.2, the components of the vectors are the conditional probability that a specific word n is produced with respect to a given topic i.

Table 5.2 Example of a Topic Vector

	word_1	word_2	word_3	...	word_n
topic_i	0.34	0.56	0.42	...	$P(w_n\|\text{topic}_i)$

Once we have a topic vector for all the possible topics, we can build a topic probability matrix consisting of all the words and possible topics. Topic vectors are obtained by training a topic model on a given corpus in a similar fashion as described for word embeddings in Chapter 4. One of the most popular approaches to train a topic model is the Latent Dirichlet Allocation (LDA) method that was first developed by Blei et al. (2003).[121] Starting from a list of documents, the idea behind LDA is to determine which words belong to a particular topic. This is done first by randomly assigning a topic to each word and then going through

[121] See Kherwa & Bansal (2018) for a comprehensive review of different approaches to topic modeling.

each document and computing the probability that a given topic t is present in a document given the proportion of words that are assigned to the topic t in this document: P(topic|document). From this, we can compute the conditional probability that a word is associated with a particular topic: P(word|topic). The topic for which this conditional probability is the largest will then be re-assigned to this word, and the process goes on for every word in all the documents. Basic LDA approaches treat every document as a list of strings, and the grammatical role of the words is not taken into account. In this thesis, I use a topic model trained on the enwiki-20170220 corpus with the LDA module of Gensim in python.

It is essential to note that the topic model' obtained from topic modeling is not the same as what I described before as the topic model as a contextual representation. The topic model is a probability distribution containing all the activated topics by an utterance at time t. These two are linked because we need to have a topic model', i.e., a topic probability matrix, to compute a representational topic model as the sum of all potential topics relevant given a specific linguistic input.

To derive the representation of a topic model, I transpose the topic probability matrix in terms of the probability of having a specific topic given a series of words. One way to do so is to sum the probability for all the topics given a particular word. When we encounter two words, e.g., *red* and *car*, these words contribute to the formation of a representation of a topic model.

Table 5.3 Transpose of the Topic probability Matrix

	topic_1	topic_2	topic_3
red	0.34	0.89	0.22
car	0.56	0.20	0.10

In Table 5.3 we see that when *red* is uttered the probability P(topic|*red*) is 0.34 for topic_1 and 0.56 for topic_2. If we also consider *car*, the conditional probability P(topic_1|*red*,*car*) becomes 0.45, i.e. the average of the conditional probabilities for both words. To obtain the representational topic model, we have to consider all the conditional probabilities for all the words in the utterance. This way, we end up with a probability distribution where the components are computed with the formula given in (132).

(132) $$P(\text{topic}_i|\text{word}_1, ...\text{word}_j) = \frac{\sum_1^j P(\text{topic}_i|\text{word}_1)}{j}$$

With (132) we can derive the representational topic model as a probability distribution using all the uttered words.

5.3.2.3 Coordination and Uncertainty

One aspect of coordination and communication I completely omitted thus far when I presented the bottom-up coordination for the situation and the topic models is the uncertain nature of the derivation process. This uncertainty occurs naturally because we rarely have all the information in hand before interpreting a sentence. In the previous paragraphs, we took for granted that the derivation of contextual level representations was straightforward in terms of the correspondence between the words that have been uttered and the words that are being interpreted as such by the listener.

To illustrate the importance of modeling uncertainty, we can go back to the sentence "John grabs the donkey.", where the word *donkey* could refer to an animal,

but it could also refer to an obstinate person. These different interpretations lead to very different situation models and topic models, and it is thus essential to determine which interpretation is favored by the listener.

To model how this lexical interpretation works, I use an extension of the RSA model that deals with uncertainties (Scontras et al., 2018). The Bayesian inference responsible for interpreting a linguistic expression requires the listener to consider the speaker's perspective. In other words, the probability that a state of the world s is retrieved from the utterance u will depend on the probability that the speaker chooses u to express s. In this uncertain RSA extension, we acknowledge the fact that the speaker's choice could be, by itself, uncertain. For example, if the speaker produces an utterance without fully knowing the state of the world, this uncertainty about s should then be taken into account by the listener when retrieving the meaning conveyed by the speaker.

In the uRSA model, the structure of the framework remains the same, but we have a new parameter a as in (133). This parameter a "refers to any factor that might influence the speaker's behavior, including uncertainty about the conversational topic, word meanings, background knowledge, or general discourse context" (Goodman & Frank, 2016, p.8). The pragmatic listener infers the state of the world and, at the same time, infers the parameter s. This is called a joint inference on the part of the listener because it involves two inferences derived simultaneously about the same thing (Scontras et al., 2018).

(133) $$P_{L_1}(s, a|u) \propto P_{S_1}(u|s, a) \cdot P(s) \cdot P(a)$$

This new parameter allows for the uRSA model to tackle more complicated issues like the interpretation of non-literal meaning (Kao et al., 2014b), metaphor un-

derstanding (Kao et al., 2014a), and even irony (Kao & Goodman, 2015). For our purpose, I use the uRSA to model the uncertainties regarding the lexical meaning of words, i.e., should we interpret *donkey* as an animal or a person. This uncertainty about lexical meaning is significant because words are the foundation on which both the topic and situation models are derived. If a word is interpreted differently, this means it might not refer to the same thing and, consequently, to the same word embeddings.[122]

Until now, we always have taken for granted that a word refers to a single word embedding, but following this lexical uncertainty problem, we should consider that a word might be associated with different word embeddings depending on its meaning. One way to model this uncertainty is to replace the lexical meaning correspondence function with a probabilistic correspondence function that attributes meanings using a set of lexical entry for every word.[123]

One way to model this is to use the RSA model's lexical-uncertainty extension that was presented in Bergen et al. (2014). In their model, they replaced the fixed lexicon $\mathcal{L}$ with a set of lexica Λ (Bergen et al., 2014). A fixed lexicon $\mathcal{L}$ is responsible for mapping utterances and states of the world via truth-values as in (134).

$$\mathcal{L}(u,s) = \begin{cases} 1 & \text{if } s \in [\![u]\!] \\ 0 & \text{if } s \notin [\![u]\!] \end{cases} \tag{134}$$

[122]Lexical uncertainty is undoubtedly linked with the noisy channel view of communication we discussed at the beginning of this chapter. In this view, a listener derives the most likely interpretation of a sentence or a lexical item while taking into account the noise of the signal that has to be interpreted (Gibson et al., 2013).

[123]*Every word* refers here to any polysemous words or words that could refer to at least two different things.

A set of lexical Λ is a probability distribution $P(\Lambda)$ containing a set of lexica $\mathcal{L}_i$, and it represents the uncertainty about the literal meaning of an expression. All of these lexica $\mathcal{L}_i$ are mapped differently with respect to their state of the world. This means that the choice of $\mathcal{L}$ influences the truth-value and thereby the global interpretation of the utterance. The approach developed by Bergen et al. (2014) was first used to model M-implicatures which are derived for cases where two utterances were semantically equivalent (Horn, 1984; Levinson, 2000). For those particular cases, they had to assign meanings and attribute a lexicon to a whole utterance.

In this thesis, I have been interested in representing utterances as a composition of word embeddings, and I must also be able to model uncertainty at the word level and not only at the utterance level. Interestingly enough, their model allows considering atomic utterances, i.e., utterances that comprise only one word, which means that we could use it to model the lexical uncertainty of a word expressed by a speaker.

However, their lexical uncertainty model's transposition into our approach of lexical prediction would not be so straightforward. One difficulty would be to know whether to resolve this lexical uncertainty before or after the compositional process. In other words, do we interpret the lexical meaning of a word and adjust it accordingly before composing the utterance together, or is it resolved after the whole utterance has been derived? In the former case, Bergen et al. (2014) describe this as *lexical enrichment*.

For the cases we described here about the derivation of the situation and the topic model, we should think of these models as representing the interpretation retrieved by the listener. Lexical uncertainty is certainly involved during the derivation of this interpretation, but we have no principled way to differentiate between a

compositional derivation or a post-compositional derivation at this point. From a practical point of view, it might be easier to resolve linguistic ambiguity before going through the derivation of both the representations at the contextual levels because then the derivation process would remain the same. In contrast, it is not clear how we could modify the situation and the topic models after their derivation.

Another crucial point is that if we want to consider lexical uncertainty and continue using word embeddings, we have to have multiple word embeddings for every word; otherwise, we would only have one mapping for every word. Word embeddings like word2vec and GloVe have only one word-vector for every word, which implies that they only have one lexicon $\mathcal{L}$ for every word. Therefore, using word2vec or GloVe, we could not model lexical uncertainty because we would not have a set of lexica attached to words.

A solution to this problem would be to use contextualized word embeddings like the ones trained with BERT (Devlin et al., 2018) but this would be beyond the scope of this thesis.[124] These two training algorithms produce contextualized word embeddings which means that they have different word-vector to represent that a word might have different meanings depending on the context it is used in. However, even if we use these contextualized word embeddings that allow a set of lexica to be used to a single word, it would not resolve the compositional/post-compositional issue about lexical uncertainty.

In this thesis, I chose not to integrate lexical uncertainties within the linguistic prediction approach because I assume that its effect will be limited since we only consider single utterances considered to be relatively unambiguous. However, lex-

[124]BERT, Bidirectional Encoder Representations from Transformers, is based on the network architecture introduced as the Transformer by Vaswani et al. (2017). This Transformer model relies on self-attention to learn contextual relations between words.

ical uncertainty undoubtedly plays a crucial role in other communication settings, especially during written conversations where the perceptual intake that could help disambiguate meanings is minimal.

5.3.3 Modeling Top-down Influences

The top-down influences from the topic model and the situation model are responsible for the constraints imposed on the potential linguistic prediction, as shown in Figure 5.14.

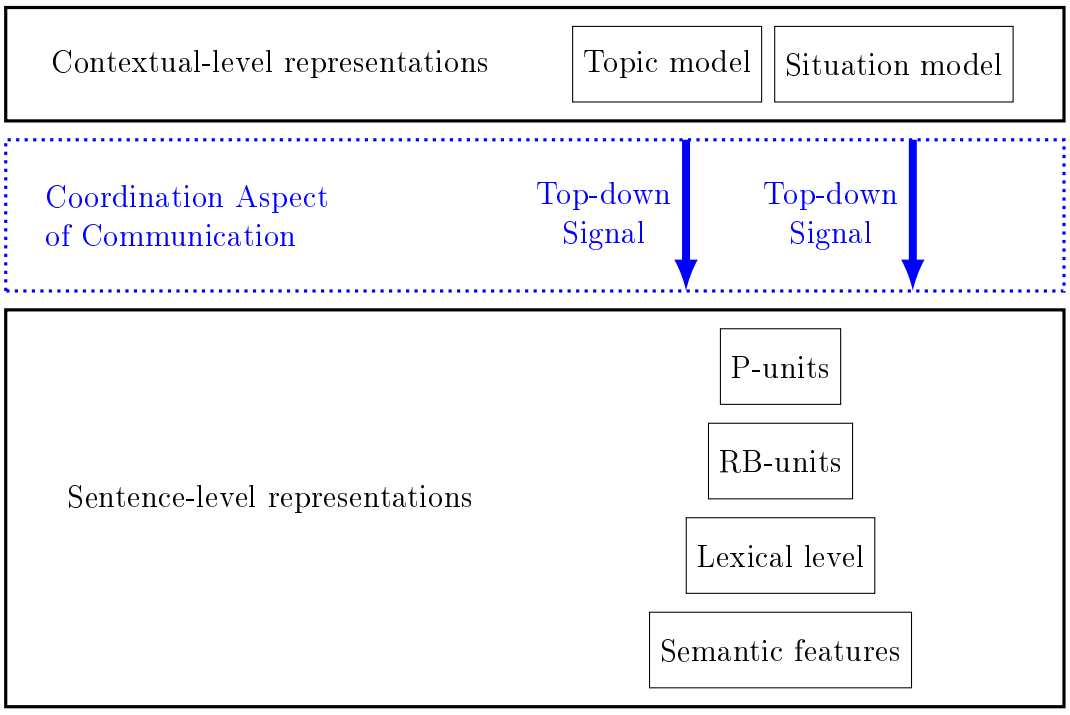


Figure 5.14 Depiction of the Top-down signals

This influence takes the form of a conditional probability imposed by these contextual-

level representations, i.e., $P(\text{input}|\text{context})$, where *context* is both the situation model and the topic model. Transposing this expression to account for linguistic prediction at the word level, we get (135).

(135) $P(\text{word}|\text{topic}, \text{situation})$

This conditional probability represents the constraints imposed upon the possible words that can be predicted according to the context. Whenever we try to predict an upcoming word, we are in such a state that this top-down influence conditions it. Under this perspective, it is not the prediction process itself that is constrained but rather the probabilistic prior. To illustrate this idea, we can go back to our example with the sentence "John plays ..." where the context of being at a chess tournament influences the probability of playing chess to compare to the probability of playing water polo.

Here, this contextual plausibility is a score conditioned on the higher-representational context that helps determine what a speaker might say or mean (Chater et al., 2006a). In terms of activation-based prediction, the top-down influences modify the upstream values in the similarity matrices for the linguistic units (P-unit, RB-unit, lexical-unit, semantic features). When taking into account the context, the conceptual space upon which is based the similarity matrices is itself updated to satisfy the conditional probability in (135).

This probabilistic approach to top-down influences models these influences to be compatible with the activation-based model of linguistic prediction described in Chapter 4. In other words, the degree of activation a of a word within the similarity matrix is determined probabilistically $P(a)$ (Heylighen, 2005) and the context conditions this probability.

From (135), we can use Bayes' rule to transform it into an expression that we can compute. Now, the fact that we consider the situation model and the topic model to be independent of each other allows us to compute the effect of both contextual models separately. It is the combination of these effects that modulate the activation level of every potential prediction.

(136) a. $P(\text{word}|\text{topic}, \text{situation}) = P(\text{word}|\text{topic}) \times P(\text{word}|\text{situation})$

b. $P(\text{word}|\text{topic}) = \frac{P(\text{word}) \times P(\text{topic}|\text{word})}{P(\text{topic})}$

c. $P(\text{word}|\text{situation}) = \frac{P(word) \times P(\text{situation}|unit)}{P(\text{topic,situation})}$

5.3.3.1 Situation Model

To compute the conditional probability that a word be predicted given the situation model in (137), we have to compute three elements: $P(word)$, $P(situation|word)$ and $P(situation)$.

(137) $P(\text{word}|\text{situation}) = \frac{P(\text{word}) \times P(\text{situation}|\text{word})}{P(\text{situation})}$

$P(word)$ is the prior probability of predicting any specific word. This prior is to be understood as the probability that a word be predicted at the time of the prediction. For example, if the linguistic prediction is derived at t_i, the prior will then be the probability computed at a previous time t_{i-1}, which is the instant that is existing right before the linguistic prediction is derived. However, because we are implementing a non-incremental model of linguistic prediction, t_{i-1} corresponds to the time when we do not have any information whatsoever about the linguistic

prediction, i.e., a time where the sentence has not been processed yet. Therefore, we suppose that we have flat priors in this model, namely that all words are equally probable.[125]

$P(situation)$ is the probability distribution that the hearer has derived a specific situation at time t_i. Doing so, we end up with a $P(situation)$ that is constituted from a probability of 1 for the situation-vector that was derived and 0 everywhere else.

The third term is $P(situation|word)$, and it corresponds to the Distributed Situation Space matrix, which consists of the correspondence between situations and words as in Table 5.4.

Table 5.4 Distributed Situation Space Similarity Matrix

	word_1	word_2	word_3	word_j
situation_i	$P(w_1\|\text{situation}_i)$	$P(w_2\|\text{situation}_i)$	$P(w_1\|\text{situation}_3)$	...
...	...	...	...	...

To compute the value of the conditional probabilities in the matrix, I use the fact that the situation model is constituted from 4 kinds of vectors concatenated together: entities, features, thematic role, and event. This means that we can compute the activation-based similarity between these components and a given word embedding separately by disjoining the situation model. This operation is readily performed by using the similarity matrices we already derived in Chapter 4. For example, when considering a word_i, we have to compute the activation-based similarity between this word and the situation model's four components.

[125]This simplification does not hold when considering incremental linguistic prediction.

To obtain the total activation-based similarity value between this word and the situation model, we can sum the activation for all of the four components.

(138) $P(\text{situation}_i|\text{word}_j) = \sum_1^4(P(\text{Components of the situation}|\text{word}_j)$

Finally, we can combine $P(\text{situation}_i|\text{word}_j)$ and $P(situation)$ to get $P(word|situation)$.

5.3.3.2 Topic Model

To compute the conditional probability that a word be predicted given the topic model in (139), we have to compute three elements: $P(word)$, $P(topic|word)$, and $P(topic)$.

(139) $P(\text{word}|\text{topic}) = \frac{P(\text{word})\times P(\text{topic}|\text{word})}{P(\text{topic})}$

Similarly, the prior is taken to be flat, which means that $P(word)$ is equal for every word. The second term $P(topic|word)$ is given by having a specific topic given a list of words. Finally, $P(topic)$ is the probability distribution represented by the topic model we derived for the truncated utterance. To sum up, the effect of the top-down influence originating from the topic model on the activation level of a word_i is given in (140).

(140) $P(\text{word}_i|\text{topic}) \propto \frac{P(\text{topic}|\text{word}_i)}{P(\text{topic})}$

For example, the probability that the word *chess* is activated given that we are at

a chess tournament is proportional to P(tournament|chess)/P(tournament), where P(tournament|chess) is the probability that the topic 'tournament' is elicited by the word *chess* and P(tournament) is the value of the probability that we are in fact as at a chess tournament.

5.4 Worked out Examples

In this section, I revisit the worked-out examples presented in Chapter 4 while taking into account the top-down influences from the topic model and the situation model. For these three cases, the situation model and the topic model were derived separately using the approach described in this chapter. In this section, I assume that the contextual-level representations do not modify the similarity values between two concepts but rather influence the concepts' pre-activation within the conceptual space. The influence coming from both the situation model and the topic model can thus be taken into account independently from the calculation we already performed in Chapter 4. Once the situation model and the topic model are derived, we can combine them with the MAPs we already calculated in the previous chapter by simply adding their contribution to the activation of the words to the activation we calculated in Chapter 4. In the following section, I display a subset of the situation model and the topic model for every worked-out example, and I show how the final MAPs change when we include the top-down influences coming from these representations.

5.4.1 Example 1: High Constraining Sentence (HCS)

For the sentence given in (141), it is the word *neck* that has the highest predictability as measured by Bloom & Fischler (1980) (96%) and by Block & Baldwin (2010) (97%).

(141) He loosened the tie around his ...

5.4.1.1 Situation Model: HCS

The first thing to do is to derive the situation model and the topic for this truncated sentence. As represented in (142), the situation model can be written as the combination of four different kinds of information: entities, features, roles, and events. The situation model of truncated sentence in (141) is represented in Figure 5.15.

(142) $\overrightarrow{\text{Situation}} = \overrightarrow{\text{Entities}} \frown \overrightarrow{\text{Features}} \frown \overrightarrow{\text{Roles}} \frown \overrightarrow{\text{Event}}$

Entities:	$\overrightarrow{\text{he}}, \overrightarrow{\text{tie}}$
Features:	$\overrightarrow{\text{male}}, \overrightarrow{\text{person}}, \overrightarrow{\text{tightness}}, \overrightarrow{\text{lessen}}, \overrightarrow{\text{clothing}}, \overrightarrow{\text{neck}}, \overrightarrow{\text{encircling}}, \overrightarrow{\text{surrounding}}$
Roles:	$(\overrightarrow{\text{loosen}} + 0.5 \times \overrightarrow{\text{he}}), (\overrightarrow{\text{tie}} + 0.5 \times \overrightarrow{\text{around}})$
Events:	$\overrightarrow{\text{loosen}}$

Figure 5.15 Representation of the situation model for the sentence in (141)

With this situation model, using similarity measures, we can compute the ac-

tivation of all the words pre-activated by this situation model. Each kind of information gives rise to a distribution of words and their associated activation. When we concatenate and add up these word-activation pairs, we get a distribution of activation levels corresponding to the effect of the situation model. A small subset of these word-activation pairs is depicted in Table 5.5.

Table 5.5 Subset of the activation coming from the situation model of (141)

tighten	4.607
stiffen	3.560
grip	2.964
cut	2.981
neck	2.276

The situation model affects the activation of words within the conceptual space regardless of their grammatical type. The syntactical constraints we discussed in Chapter 4 played a role when deriving a prediction, and they were already taken into account when deriving the MAPs at every level. However, there is no need to implement them here since we are computing the intersection of those two sets. For example, if a word is present in the situation model and absent from the MAPs we calculated in Chapter 4, it will simply not appear in the prediction. Conversely, if a word is present in the MAPs and is not present in the situation model, it will still be a valid prediction, but it will not be pre-activated by the situation model.

(143) $\text{MAP}_{\text{Chap 4}} \cap$ Situation Model

Once we intersect and superpose the situation model and the MAPs we got in

Chapter 4, we can retrieve the situation model's final prediction. A comparison between the results in Chapter 4 and those obtained when taking into account the situation model is shown in Table 5.6.

Table 5.6 First five MAP for the situation model of (141)

Chapter 4		with Situation Model	
back	0.055	neck	13.702
neck	0.052	thigh	13.499
thigh	0.051	ankles	12.089
rib_cage	0.051	wrists	12.008
legs	0.051	back	11.829

It is interesting to see that once we combine the activations from the situation model and those we already calculated for that MAPs in Chapter 4, the word *neck* now becomes the most activated word. Conversely, the word *back*, because the situation model activates other words more, is now relatively lower than before.

5.4.1.2 Topic Model: HCS

To include the top-down effect coming from the topic model, we have to determine which topics are associated with the words present in the truncated sentence of (141). This first operation is done using a word-topic probability matrix obtained from the training of a topic model.[126] To determine which topic is the most

[126]As I already mentioned, the topic model used in this chapter was trained on the enwiki-20170220 using the LDA module of Gensim. The result was a model containing 150 different topics and their associated word-topic probability.

probable, I used the words *loosened* and *tie* that were expressed in (141).[127]. As a result, I obtained that the most probable topic was the number 049, and the probability was 0.5033. All the other topics had the same probability of 0.00333 and were not included in the calculation.

From that topic, we can retrieve the words that are the most probable when discussing this topic. A subset of the resulting distribution of words is given in 5.7.

Table 5.7 Subset of the activation coming from the topic model of (141)

match	0.013
round	0.013
rank	0.011
championship	0.010
draw	0.010

Then, similarly to what we did for the situation model, we can find the intersection of this set of words with the one from the MAP we calculated in Chapter 4 and add up the contribution of the topic model. A comparison between the results in Chapter 4 and those obtained when taking into account the topic model is shown in Table 5.8.

[127] I could not used either *he* or *around* since they are not accounted for during the training because they usually bear little topical content (Wang et al., 2011)

Table 5.8 First five MAP for the topic model of (141)

Chapter 4		with Topic Model	
back	0.055	thigh	9.326
neck	0.052	forearm	8.560
thigh	0.051	elbow	7.365
rib_cage	0.051	neck	6.876
legs	0.051	rib_cage	6.875

5.4.1.3 Deriving the Linguistic Prediction: HCS

Finally, we can add the contribution from the situation model and from the topic model to obtain the linguistic prediction.

Table 5.9 First five MAP for the sentence (141)

neck	13.702
thigh	13.499
ankles	12.089
wrists	12.008
back	11.829

5.4.2 Example 2: Low Constraining Sentence (LCS)

For the sentence given in(144), the cloze scores are the following (Bloom & Fischler, 1980): *hands* (49%) , *glove* (32%), *mitt* (8%), *teeth* (4%).

(144) He caught the ball with his ...

5.4.2.1 Situation Model: LCS

The first thing to do is to derive the situation model and the topic for this truncated sentence. The situation model of truncated sentence in (144) is represented in Figure 5.16.

(145) $\overrightarrow{\text{Situation}} = \overrightarrow{\text{Entities}} \frown \overrightarrow{\text{Features}} \frown \overrightarrow{\text{Roles}} \frown \overrightarrow{\text{Event}}$

Entities:	$\overrightarrow{\text{he}}, \overrightarrow{\text{ball}}$
Features:	$\overrightarrow{\text{male}}, \overrightarrow{\text{person}}, \overrightarrow{\text{receive}}, \overrightarrow{\text{object}}, \overrightarrow{\text{thrown}}, \overrightarrow{\text{round}}, \overrightarrow{\text{game}}, \overrightarrow{\text{instrument}}$
Roles:	$(\overrightarrow{\text{catches}} + 0.5 \times \overrightarrow{\text{he}}), (\overrightarrow{\text{catches}} + 0.5 \times \overrightarrow{\text{ball}})$
Events:	$\overrightarrow{\text{catches}}, \overrightarrow{\text{catcher}}$

Figure 5.16 Representation of the situation model for the sentence in (144)

From this situation model, we can compute the activation of all the words that are pre-activated by it. When we concatenate and add up these word-activation pairs, we get a distribution of activation levels corresponding to the effect of the situation model. A small subset of these word-activation pairs is depicted in Table 5.10.

Once we intersect and superpose the situation model and the MAPs we got in Chapter 4, we can retrieve the situation model's final prediction. A comparison between the results in Chapter 4 and those obtained when taking into account the topic model is shown in Table 5.11.

Table 5.10 Subset of the activation coming from the situation model of (144)

guy	4.039
bat	3.815
him	3.715
throw	3.544
hands	1.385

Table 5.11 First five MAP for the situation model of (144)

Chapter 4		with Situation Model	
hands	0.164	hands	10.046
glove	0.118	glove	8.765
mitt	0.078	mitt	5.543
socks	0.077	stick	3.127
toes	0.040	knee	2.742

5.4.2.2 Topic Model: LCS

To determine which topic is the most probable, I used the words *caught* and *ball* that were expressed in (144). As a result, I obtained that the most probable topic was the one with the number 120 and the probability was 0.669. All the other topics had the same probability of 0.00222 and were not included in the calculation. From that topic, we can retrieve the words that are the most probable when discussing this topic. A subset of the resulting distribution of words is given in 5.12.

Then, similarly to what we did for the situation model, we can find the intersec-

Table 5.12 Subset of the activation coming from the topic model of (144)

he	0.049
but	0.047
hand	0.042
glove	0.040
fingers	0.039

tion of this set of words with the one from the MAPs we calculated in Chapter 4 and add up the contribution of the topic model. A comparison between the results in Chapter 4 and those obtained when taking into account the topic model is shown in Table 5.13.

Table 5.13 First five MAP for the topic model of (144)

Chapter 4		with Topic Model	
hands	0.164	hands	0.164
glove	0.118	glove	0.118
mitt	0.078	mitt	0.078
socks	0.077	fingers	0.077
toes	0.040	socks	0.040

5.4.2.3 Deriving the Linguistic Prediction: LCS

Finally, we can add the contribution from the situation model and from the topic model to obtain the linguistic prediction.

Table 5.14 First five MAP for the sentence (144)

hands	9.239
glove	8.765
mitt	5.543
stick	3.856
toes	1.832

In 5.14, the order of the first three MAPs is still the same as the one obtained by (Bloom & Fischler, 1980), but the word *teeth* is still not present in the list of MAPs.

5.4.3 Example 3: The influence of the context (IoC)

In this example taken from Nieuwland & Van Berkum (2006), the association between the meaning of *peanut* and the possible continuation is modified by the use of a specific context (146) where the peanut is attributed anthropomorphic characteristics. In Chapter 4, we retrieved the linguistic prediction by looking only at the truncated sentence (147). Here, we derive the situation model and the topic model of this preceding context (146), and we integrate their contribution into the model of linguistic prediction.

(146) [Full context] A woman saw a dancing peanut who had a big smile on his face. The peanut was singing about a girl he had just met. And judging from the song, the peanut was totally crazy about her. The woman thought it was really cute to see the peanut singing and dancing

like that.

(147) The peanut was [salted/in love].

5.4.3.1 Situation Model: IoC

The situation model of the preceding context (146) is given in Figure 5.17.

Entities:	$2 \times \overrightarrow{\text{woman}}, 3 \times \overrightarrow{\text{peanut}}, \overrightarrow{\text{face}}, \overrightarrow{\text{girl}}, \overrightarrow{\text{he}}, \overrightarrow{\text{her}}, \overrightarrow{\text{song}}$
Features:	$4 \times \overrightarrow{\text{female}}, 5 \times \overrightarrow{\text{person}}, \overrightarrow{\text{body}}, 6 \times \overrightarrow{\text{human}}, \overrightarrow{\text{male}}, 3 \times \overrightarrow{\text{legume}}, 3 \times \overrightarrow{\text{edible}}, \overrightarrow{\text{music}}, \overrightarrow{\text{words}}$
Roles:	$(\overrightarrow{\text{saw}} + 0.5 \times \overrightarrow{\text{woman}}), (\overrightarrow{\text{saw}} + 0.5 \times (\overrightarrow{\text{dancing}} + 0.5 \times \overrightarrow{\text{peanut}})), (\overrightarrow{\text{smiling}} + 0.5 \times \overrightarrow{\text{face}}), 2 \times (\overrightarrow{\text{singing}} + 0.5 \times \overrightarrow{\text{peanut}}), (\overrightarrow{\text{met}} + 0.5 \times \overrightarrow{\text{he}}), (\overrightarrow{\text{met}} + 0.5 \times \overrightarrow{\text{girl}}), (\overrightarrow{\text{crazy}} + 0.5 \times \overrightarrow{\text{peanut}}), (\overrightarrow{\text{see}} + 0.5 \times \overrightarrow{\text{peanut}}), (\overrightarrow{\text{see}} + 0.5 \times \overrightarrow{\text{peanut}}), (\overrightarrow{\text{dancing}} + 0.5 \times \overrightarrow{\text{peanut}}), (\overrightarrow{\text{thought}} + 0.5 \times \overrightarrow{\text{woman}})$
Events:	$\overrightarrow{\text{saw}}, 2 \times \overrightarrow{\text{dancing}}, \overrightarrow{\text{smiling}}, 2 \times \overrightarrow{\text{singing}}, \overrightarrow{\text{met}}, \overrightarrow{\text{crazy}}, \overrightarrow{\text{thought}}, \overrightarrow{\text{see}}$

Figure 5.17 Representation of the situation model for the sentence in (147)

The transcription of the context in (146) into the situation model is not straightforward. First of all, some predicates like *having-a-smile* are not present in our corpus, so we had to represent them using other predicates, *smiling* in this case. For the same reason, we also simplified the representations of some roles. I used the predicate *saw* for both the woman and the peanut even though the woman is an agent and the peanut a theme. Also, I do not consider the adverbial clauses even though they contribute to the meaning expressed by a sentence. For example, the clause *judging from the song* is not represented in this situation model. In this

thesis, I have tried my best to remain faithful to the principle behind the LISA and DORA representational approaches, but it is clear that the transposition of longer and more complex propositions into word embeddings forces us to make some simplification. It would be relevant to develop a more formal and systematic transposition mechanism in the near future, but this would be outside the scope of this thesis.

From this situation model, we can compute the activation of all the words that are pre-activated by it. When we concatenate and add up these word-activation pairs, we get a distribution of activation levels corresponding to the effect of the situation model. A small subset of these word-activation pairs is depicted in Table 5.15.

Table 5.15 Subset of the activation coming from the situation model of (147)

man	12.956
girl	12.551
lady	12.330
love	7.950
salted	1.549

In Table 5.15, it is interesting to note that *love* significantly more activated than *salted*.

Once we intersect and superpose the situation model and the MAPs we got in Chapter 4, we can retrieve the situation model's final prediction. A comparison between the results in Chapter 4 and those obtained when taking into account the topic model is shown in Table 5.16.

Even though they are not part of the five most activated linguistic prediction, the words *love*, *lovely*, *loves* all have a quite high activation, i.e., respectively 0.06,

Table 5.16 First five MAP for the situation model of (147)

Chapter 4		with Situation Model	
milk	0.028	chocolate	0.009
wheat	0.028	drink	0.009
dairy	0.027	bird	0.008
meat	0.027	sweet	0.006
grain	0.027	meat	0.006

0.05, and 0.04. Most importantly, these three words are all more activated by the situation model than the words *salty* (0.002) and *salted* (0.001).

5.4.3.2 Topic Model: IoC

As explained before, to determine the most probable topic, I used all the words in (146), except the non-topical words like *he* or *she*. As a result, I obtained that the most probable topics were the ones with numbers 4, 10, and 12. Their respective probability of occurrence is 0.199, 0.510, and 0.250. All the other topics were not included in the calculation. From these topics, we can retrieve the words that are the most probable when discussing these topics. A subset of the resulting distribution of words is given in 5.17.

Here, the words *love* and the word *loves* are activated by the topic model, whereas the word *salted* is not activated at all.

Then, similarly to what we did for the situation model, we can find the intersection of this set of words with the one from the MAP we calculated in Chapter 4 and

Table 5.17 Subset of the activation coming from the topic model of (147)

rio	0.00109
brazil	0.00105
hotel	0.00089
restaurant	0.00065
love	0.00037

add up the contribution of the topic model. A comparison between the results in Chapter 4 and those obtained when taking into account the topic model is shown in Table 5.18.

Table 5.18 First five MAP for the topic model of (147)

Chapter 4		with Topic Model	
milk	0.028	milk	0.0205
wheat	0.028	wheat	0.0203
dairy	0.027	dairy	0.0194
meat	0.027	meat	0.0194
grain	0.027	rice	0.0192

When taking into account the topic model, the activations for the words *love* and *salted* is equal at 0.0007.

5.4.3.3 Deriving the Linguistic Prediction: IoC

Finally, we can add the contribution from the situation model and from the topic model to obtain the linguistic prediction.

Table 5.19 First five MAP for the sentence (147)

cook	0.011
chocolate	0.0091
drink	0.0086
eat	0.0085
bird	0.0078

The interesting results here are not the five most activated prediction, but the fact that when we take into account the situation model and the topic model, we get that *love* is now more highly activated at 0.005 than the word *salted* which is activated at 0.001. This means that the situation model and the topic models have changed the expectation of encountering the word *love*. It was absent from the distribution of linguistic prediction in Chapter 4, and it now has become more activated, or more highly expected, than the word *salted* or the word *salt* (0.003).

5.5 Discussion: A Cognitive Architecture for Linguistic Prediction

Chapter 2 presented the cognitive architecture for linguistic processing proposed by Baggio (2018). I now compare this cognitive architecture with the multi-layered representation of linguistic prediction presented throughout this chapter. This multi-layered representation corresponds to the hearer's different representational levels needed to predict the upcoming word, and it is illustrated in Figure 5.18.

Let me recall that Baggio (2018) described three systems that were involved in

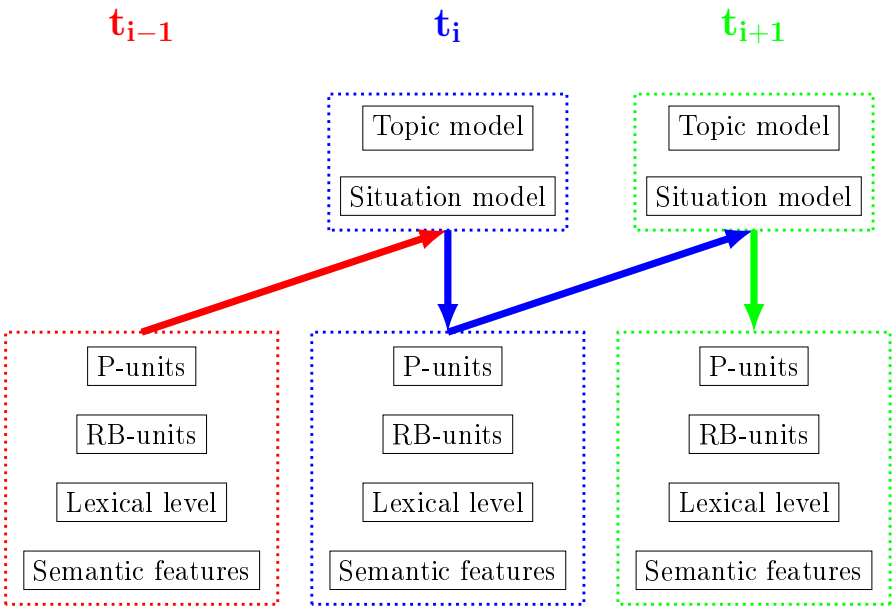


Figure 5.18 Representational processing structure involved in linguistic prediction

linguistic processing. The R-system is responsible for mapping lexical items into relational structures and is triggered every time a new word is processed. The I-system is responsible for interpreting these relational structures and their transposition into a mental model of the input, and it is triggered at every new referring expression. Finally, the E-system is linked with the coordination aspect of communication, and it treats whole signals or whole communicative actions. The cognitive architecture from Baggio (2018) is depicted in Figure 5.19 (Baggio, 2018, adapted from p.186).

To find the correspondence between Figure 5.18 and Figure 5.19, i.e., to transpose the different characteristics of my multi-layered approach of linguistic prediction into the cognitive architecture from Baggio (2018), we have to look more closely at every stage of processing. Baggio (2018) distinguished two main processing paths:

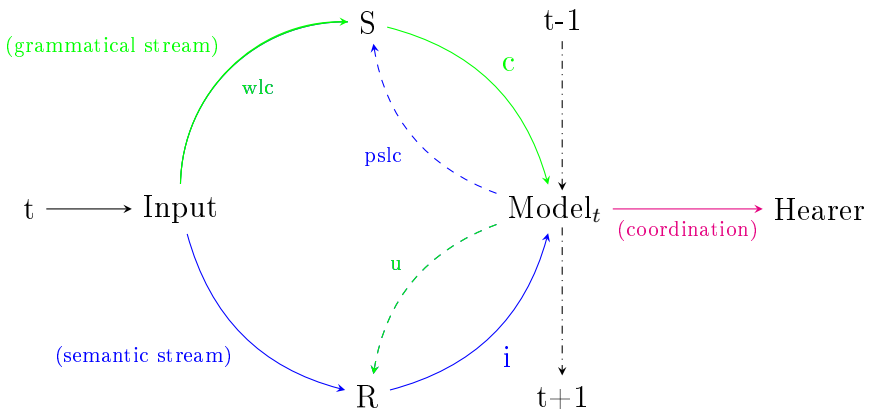


Figure 5.19 Cognitive architecture for Linguistic Processing

the grammatical path, which is related to the syntactic processing we discussed in Chapter 2, and the semantic path, which has to do with the semantic processing stream.

It is important to note that these two streams co-exist but that their relative contribution varies from word to word (Baggio, 2018). For example, the grammatical stream might have a greater contribution for words where the lexical meaning is unknown. In the linguistic prediction model presented here, this word-to-word relative contribution was not considered since our model is not incremental. However, I did acknowledge that semantics usually had precedence over syntax by focussing mainly on the semantic stream.

In Figure 5.19, the grammatical stream is represented by the following path: $I \rightarrow S \rightarrow M(\rightarrow R)$. The input word I is first grammatically analyzed with respect to word-level constraints (i.e., *wlc*) to transform it from a token into a grammatical type representation at S. From there, the model M is derived using compositional operations. In my linguistic prediction approach, I did not explicitly discuss the specifics involved in the syntactic stream in Chapter 4. However, it

was included in the discussion about compositional units because determining the P-unit and the RB-units of a composition presupposes a grammatical analysis of every component of this composition. In other words, to retrieve the four levels of compositional units, we had to consider this *wlc* process implicitly. Within this grammatical stream, the final step from M to R is the optional updating of the stored representations of semantic types at R. Besides, this constraint imposed by the grammatical stream via the process u also indirectly influences the prediction concerning the semantic types or the upcoming word's semantic features.

The semantic stream is represented by the arrows going from $I \rightarrow R \rightarrow M(\rightarrow S)$. Here, the token input I is bound with the previous linguistic context as a relational structure of semantic types. This relational structure R is then interpreted (i) and included within the model M. Once M is updated, it might impose phrase and sentence level constraints (*pslc*) on the grammatical types represented at S. For example, an indefinite NP might have a higher probability than a definite NP in a linguistic context where the referent of this definite NP is not explicitly expressed. When it comes to linguistic prediction, the step *pslc* was described in Chapter 4 as a constraining mechanism that influenced by limiting the possible grammatical types of possible continuation, i.e., the grammatical type of the upcoming word has to bear a specific grammatical type for the complete sentence to compose. To illustrate this, we could compare the three contexts and the possible continuation in (148).

(148) a. Peter has recently visited a speaker in Munich.
He said that the/a speaker had been very nice.

b. Peter has recently visited a lecture in Munich.
He said that the/a speaker had been very nice.

c. Peter has recently met Hannah in Munich.

He said that the/a speaker had been very nice.

If the chosen word is *the* it will form a definite NP whereas it constitutes an indefinite NP if *a* is used. When there is no clear referents available from the linguistic context as in (148-b) and (148-c), *a* should thus be a more probable continuation than *the*. In fact, larger P600 effect were elicited for those two cases compared with (148-a) (Baggio, 2018).

Once again, in terms of our linguistic prediction model, the processes r and i were not explicitly described, and I decided to directly consider the sentence-level units (P-unit, RB-units, lexical units, and semantic features units) as given for a particular sentence. The multi-layered representations approach of linguistic prediction presented in this chapter depicts the structural interaction between these different representational levels, but it does not consider how these sentence-level units are derived in the first place.

Going back to Figure 5.19, it is tantamount to say that the model of linguistic prediction originates at M in the figure. However, the model we are considering as a starting point is solely constituted from the sentence-level units. In other words, the grammatical and semantic streams are responsible for the derivation of the minimal model M that consists of the sentence-level units associated with a given sentence.

The contextual-level representations are higher on the representational scale than this minimal model, which means they should be distinguished from the model M from Baggio (2018). This implies that the situation model and the topic model, which are conceptual constructions derived from this simple sentential model M, should be represented at a different level in this architecture. Putting all of this together, we can represent my architectural model in terms of the three systems

described by Baggio (2018).

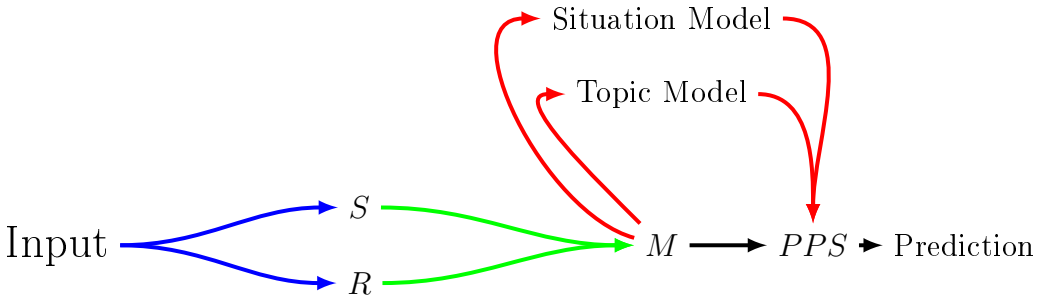


Figure 5.20 Representational structure of this model of linguistic prediction

Figure 5.20 shows that the R-system ($I \rightarrow R$) lies outside the scope of our model of linguistic prediction. This is not a surprise because the model's primary inputs are word embeddings that correspond to semantic types or grammatical types. In the I-system, the composition of different word embeddings gives rise to the model M, which contains the four sentence-level units derived using both the semantic types from the word embeddings and the grammatical types associated with these word embeddings. The core of this linguistic prediction model sits within the E-system, which is responsible for the coordination aspect of communication. This coordination between the speaker and the hearer allows us to derive the situation model and the topic model from the compositional model M.

The right-hand side of Figure 5.20 represents the predictive process itself, and it is not part of any processing system per se, but it could well be integrated within the E-system because it is also related to the coordination aspect of communication. I already argued that both the bottom-up signals and the top-down influences involved coordination: the first one is coordination between a speaker and a hearer, and the second one is coordination between states of the world.

It would therefore be natural to include these top-down influences into the E-system. However, when it comes to linguistic prediction, it is not the minimal model M that is constrained but the activation-based similarity network derived from this model. In other words, it is not the representations from the model that are directly affected by the contextual level but the level of activation of other lexical items related to these representations. This is why, in Figure 5.20, the pre-prediction state (PPS) is represented as an activation-based network obtained from the model M and influenced both by the situation model and the topic model. The linguistic prediction is then derived from this PPS.

In Figure 5.21, I illustrated the general architecture of the model of linguistic prediction discussed in this thesis in terms of the processing architecture from Baggio (2018). To the general architecture that we already discussed regarding the composition of meaning, I added a higher representational level to refer to the coordination aspect we described in this Chapter. According to Figure 5.21, both the Topic model and the Situation model would be considered being part of the E-system represented in magenta which may influences the way the next input is processed.

When we discussed the uncertainty that is inherently involved in the coordination aspect of communication, I mentioned that the resolution of such lexical uncertainty was realized either before or after the composition. In terms of the three processing systems, this implies that the I-system could transform types into referents before or after the derivation of the minimal model M. Even though Baggio (2018) stipulates that the E-system also contains the I-system, he also argues that the E-system operates from the I-system outputs as depicted in Figure 5.21 since both the Topic and the Situation Models are derived from the Model_t itself. In this thesis, I am agnostic regarding the pre- versus post- compositional resolution of these lexical ambiguities, but it should be noted that the pre-compositional view

would imply a direct link between of the E-system on the I-system, i.e. instead of being derived from the Model$_t$, the Topic and the Situation Models would be derived directly from R in Figure 5.21.[128]

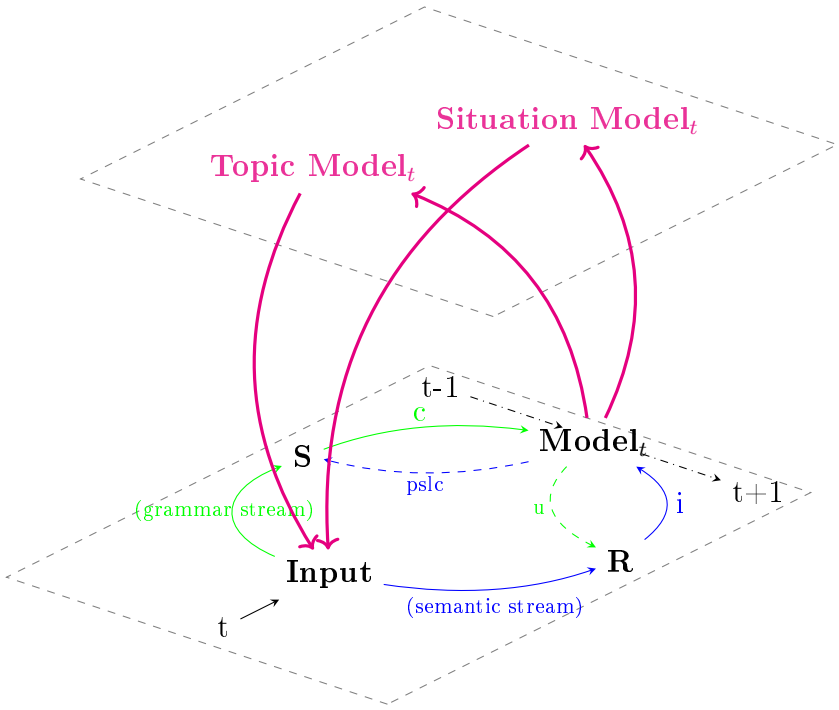


Figure 5.21 General architecture of this model of linguistic prediction

Despite its apparent complexity, the existence of this representational processing structure does not contradict the view that language processing is often based on shallow processing. (Ferreira & Swets, 2002; Ferreira & Lowder, 2016), which stipulates that when processing linguistic inputs, we often end up with representations that are not fully parsed or chunked (Kaiser, 2013).

The present model is also in line with this view, generally called the "now-or-never

[128]Figure 5.21 is presented to make a parallel between the general framework of linguistic prediction presented in this thesis, and it was beyond the scope of this thesis to develop an exhaustive representation of linguistic processing that would include the coordination aspect of communication.

bottleneck" (Christiansen & Chater, 2016), which bears some similarities with the "good-enough processing" view (Ferreira & Lowder, 2016). Basically, according to the "now-or-never bottleneck" view, incoming information has to be treated as fast as possible to keep some cognitive resources available to treat any new incoming information. This bottleneck is generally described as a processing constraint on perception and action. One way to save cognitive space from processing much incoming information at the same time is to chunk it and represent it at different representational levels (Christiansen & Chater, 2016).

This bottleneck is easy to understand when we think of language processing because when someone is talking, we have to process a vast amount of information in a short amount of time. After all, if we cannot keep up, newer information rapidly overwrites the older one. To avoid this, linguistic information is processed and chunked rapidly to form a linguistic structure which is then processed at a higher level of representation (Christiansen & Chater, 2016). The fact that the contextual level is part of this representational structure implies that it must be taken into account during the chunking process, but, fortunately, the spreading of activation happening at different representational levels is a relatively shallow and effortless cognitive process.

5.6 Summary

In this chapter, I develop a linguistic prediction model that underlines the importance of the contextual-level representations and the top-down influences that result from them. Also, I argued that the multi-layered representational processing structure shapes how these linguistic predictions are derived. Doing so, I am

supporting the idea that language-centric explanations and contextual representations at the discourse-level (Hasson et al., 2018) have to be taken into account if we aim to have a deeper understanding of linguistic prediction.

I used word embeddings as the primary input for the model and integrated these into a representational structure that processed the information in a certain way to constrain the derivation of the linguistic prediction. These word embeddings carry primordial statistical information about how words are used in natural language (Christiansen & Chater, 2016), but it is the combination of this statistical information and the proposed processing structure that gives out a prediction. In other words, when developing a generative model of linguistic prediction, we not only need the statistics of use, but we have to structure these statistics to understand the causal matrix responsible for linguistic prediction (Clark, 2013), i.e., we have to think of representation structures as a processing mechanism (Christiansen & Chater, 2016).

In the last section of this chapter, I have revisited the three worked-out examples discussed in Chapter 4. In the case of the High Constraining Sentence, we now have the word *neck* as the most activated prediction, and this corresponds to the empirical result obtained by (Bloom & Fischler, 1980). In the Low Constraining case, as was the case before, the ordering for activations for the words *hands*, *glove*, and *mitt* exactly corresponds to the ordering for the cloze scores obtains by Bloom & Fischler (1980). Finally, in the case where we had to compute the relative activation of the word *love* and the word *salted*, the model did activate the word *love* more than the word *salted* which points towards a strong contribution of the contextual representations. Once again, these results support the empirical adequacy of this model.

CHAPTER VI

DISCUSSION

In the previous chapters, I have presented a theoretically oriented model of linguistic prediction where, given an input of words and sentences, the model gives a ranked list of possibilities for the next word. This model derives the most likely continuations for a given sentence from the meaning expressed in that truncated sentence and, for the first time, the semantics of prior discourse. To derive the linguistic predictions, I used general pre-trained word embeddings, rather than a closed world with pre-determined world-states like most connectionist approaches (Schuster et al., 2020). In this model, word embeddings were used to create different levels of representations like semantic features level, lexical level, RB level, and P level. I then developed a theoretical model to generate the contribution from the contextual level representations, which were also derived using word embeddings. The contribution from these contextual models was then taken into account when predicting the upcoming word. I also presented three worked-out examples illustrating how this model of linguistic prediction could be implemented, and I showed the model made the correct predictions for three kinds of cases. This combination of the contributions coming from the sentence and contextual levels achieved empirical adequacy in that the prediction matched the attested experimental results.

In this thesis, my motivations were to present a cognitively and linguistically sound approach to the influence of contextual-level representations. The model of linguistic prediction presented in this thesis represents the first step towards a better understanding of the role of pragmatics within linguistic prediction and linguistic processing more generally. Out of necessity, this thesis has prioritized the presentation of the tools and motivations to build this highly transdisciplinary approach. As such, some simplifications were posited along the way.

As we continue to develop a more sophisticated implementation of this model of pragmatic processing in future work, there is a number of other considerations to improve the existing model, and those will be explored in section 6.1. The insights present in this model could also be compared with other frameworks, and this is discussed in section 6.2. Finally, in section 6.3, I present a different and more mathematical way to model conceptual spaces.

6.1 Extensions to the Model

In this section, I discuss four kinds of possible extensions. The first has to do with the parametrization of the model. In this thesis, I present a base model which could later be fine-tuned. This parametrization will not change the results presented in the previous chapters, but it could widen the scope of application to model more precisely other kinds of complex examples. For the second discussion, I revisit the incremental nature of linguistic processing discussed in the previous chapters. It is essential to mention that the tools and the approaches were chosen so that the model could be incremental, even though it is not currently implemented in an incremental way. Another aspect that I did not tackle is

that cognitive processes are costly, meaning we have to consider processing costs when modeling linguistic prediction. Finally, I return to an aspect that I briefly discussed in Chapter 1: the individual variability in the cloze scores.[129]

6.1.1 Parametrization and the Vanilla Model

The worked-out examples presented in Chapter 4 and Chapter 5 were derived using a "vanilla" model of linguistic prediction, which means that some parameters were pre-determined or not present for simplicity. As we improve the model, we are able to modify some parameters without modifying the model's core structure. In this subsection, I look at three kinds of possible parametrization that could be more thoroughly examined in future model implementations.

6.1.1.1 Threshold of activation

A threshold of activation is a score below which a concept is considered not to be activated. In the vanilla model presented here, we did not determine the threshold of activation empirically. Instead, we selected all the possible candidates every time, even though it might not be cognitively realistic because it is unlikely that a participant always activates all words semantically related to another word.

[129]The fact there are several possible values for the different parameters used in the implementation of the model that was presented in the preceding chapters should not be seen as a hindrance to the development of this general theory of next-word prediction. In this Chapter, I discuss different avenues that could be explored or improved without questioning the basic assumptions behind this theoretical framework.

To have a more cognitively realistic model, we could add a parameter for the activation threshold, below which words will not be activated. This way, when a word is encountered in the sentence, it only activates the most highly similar words.

6.1.1.2 Weight of Contribution

In Chapter 4, the four sentence-level representations (SF-units, Lexical-units, RB-units, P-unit) contribute equally to the derivation of a linguistic prediction. However, this might not be the case for all situations. Yun et al. (2012) found that semantic similarity had a more substantial effect when the linguistic information expressed in the sentence was less constraining, and it had a weaker effect when the linguistic context was more constraining. A more constraining linguistic context leads to faster processing because the predicted word is easier to activate above a certain threshold of activation (Staub et al., 2015). The constraints coming from a sentence are linked to the number of possible endings, and these endings are themselves linked to the relative entropy over all the possible words that could complete this sentence (Kuperberg, 2016). In other words, the difference in activation between the most and the least activated concept is greater for a high-constraining context than for a less constraining context (Staub et al., 2015).

In Yun et al. (2012), they only considered one level of representation, namely the lexical level, and they compared the contribution from similarity with that from predictability. My model is different because predictability is modeled from indirect measures of similarity. In terms of the four sentence-level representations,

this result can be interpreted as giving more weight to the lexical level in a less constraining context because the contribution from the other sentence-level representations is not enough. To model the difference between strongly constraining and weakly constraining, we could add additional weight to the four different sentence-level representations' relative contributions.

6.1.1.3 Local and Broader Context

Finally, it is essential to discuss the difference between the local and the broader context. We can illustrate this difference by looking at a sentence like (149).

(149) *The day was breezy, so the boy went to the park to fly...*

In a sentence like this, the local context might be the particle *to fly...* or even *to the park to fly...*, while the broader context is the first part of the sentence, i.e., *The day was breezy.* In (149), the local context plays a role in the derivation of the linguistic prediction, i.e., *kite*. However, in this particular case, the broader context also seems to play a role since *breezy*, and *boy* also contribute to constraining the number of possible candidates for the continuation.

As described by Smith & Levy (2013), many empirical results have shown that local context models usually outperform broader context models when looking at the accuracy of responses. However, this is not to say that the broader context does not contribute to the prediction, but it supports the idea that locality effects have a more considerable impact on sentence processing. One recent model that is in line with this statement about the local context is the lossy-context surprisal

model from Futrell et al. (2020); Futrell (2019); Futrell & Levy (2017). According to their approach, a linguistic prediction is derived using a noisy memory representation, i.e, a probabilistic representation of the information coming from previous words or from prior discourse (Futrell et al., 2020)

They decompose the surprisal into two parts: an unconditional surprisal and a Pointwise Mutual Information (PMI) component (Futrell & Levy, 2017). In (150), the first term represents the surprisal of having the word w_i without any linguistic context, and the PMI is an information-theoretic measure of the strength of association (Futrell et al., 2019) between two linguistic entities. PMI can also be viewed as comparing the number of shared bits between two representations (Futrell et al., 2020).

(150)

$$D_{surprisal}(w_i|Context) \propto -log[p(w_i|Context)] = \\ - log[p(w_i)] - pmi(w_i|Context)$$

The idea here would be to integrate their lossy-context approach directly into the model. Even though theirs is a surprisal-based model while mine is predictability-based, we could still take this lossy-context into account when measuring the similarity between words at the lexical level. If we assume a lossy-context representation, then it implies that we also have a principle of information locality (Gibson et al., 2019). In other words, if the representation of the context becomes noisier as we go back in time, it means that the local context must have a more substantial contribution when predicting the next word.

In Futrell et al. (2020), this principle of information locality is defined as the

erasure noise that reduces the PMI linearly as the distance increases. To illustrate this, we go back to the sentence about the *kite*, and we could add two weights that are normalized according to their distance to the predicted word, e.g., *day* has the smallest weight, and *fly* has the highest one, as we can see in (151).

(151) *The [0.2 day] was breezy so the boy went to the park to [0.8 fly]....*

In the present model, the erasure probability e_d that increases monotonically with the distance d could be transposed as a normalized weight that increases linearly.

6.1.2 Incrementality

As discussed in Chapter 3, incrementality is a desideratum when developing a model of linguistic prediction. However, a non-incremental model is easier to implement as a base model, and, in this thesis, my focus was on describing the pragmatic stream's structural processing view. It is essential to mention that the model presented in this thesis is inherently compatible with an incremental processing view, even though it has not been integrated into the model yet.

In this model, I consider that a linguistic prediction is performed at a given point in time and that, when trying to determine the upcoming word of a truncated sentence, we are using all the information accessible to us at this time. Thus, the pre-predicting processes that consist of collecting the information used at the predicting stage and the prediction process are separate. The prediction phase is the transition between the representation of world knowledge, consisting of all the representations available to the hearer at that time and the predicted word.

In contrast, the pre-prediction phase consists of the derivation of those representations from the given input as in Figure 6.1.

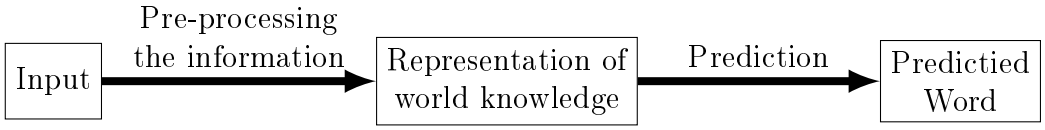


Figure 6.1 Depiction of a Linguistic Prediction in Terms of Processing Phases

The model presented in this thesis allows the incremental build-up of the representations of world knowledge, and it is thus compatible with an incremental view of this pre-predicting phase. However, when considering incrementality, we must make sure that the processes involved in the derivation of these representations are also incremental.

Concerning the syntactic and semantic processes involved in linguistic processing, we saw in Chapter 3 that incrementality is not a problem. I presented linguistic approaches that are compatible with incrementality, such as Combinatory Categorial Grammar (CCG) (Steedman, 1996, 1999), Discourse Representation Theory (DRT) (Kamp et al., 2011; Kamp, 2008, 1995), and Relevance Theory (Sperber & Wilson, 1995) which acknowledged the incremental aspect of pragmatics during the back-and-forth derivation process of the implicature and the explicature. I also discussed the incremental nature of game-theoretic pragmatics because, in a 'syntactic game' (Skyrms, 2010), we can always divide the signal into a sum of sub-signals, where each sub-signal is like a word within a complete sentence. Viewing every input as a series of sub-inputs means that this time t at which an interpreter predicts the upcoming word could well be divided as a series of transitory moments.

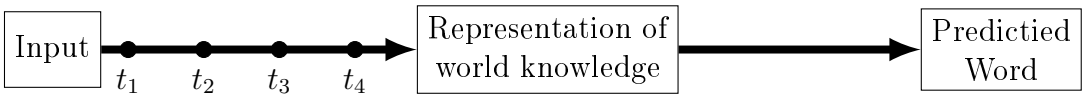


Figure 6.2 Incremental Processing in Terms of Different Temporal Moments

In Figure 6.2, the time before which the prediction is derived is divided into a series of moments t_i, and each of these moments is a moment where the representation of world knowledge is derived from the information available at that moment. In other words, there should be no principled distinction between the processes involved at each moment of the pre-predicting phase of the sentence being developed in (152).

(152) a. John went to the ...

b. John went to the park to ...

c. John went to the park to fly a ...

Every time a participant has to predict the upcoming word, this prediction is based on their representation of world knowledge at the time of this prediction. This is not to say that the information available to the interpreter at all these moments is the same, but rather that the processes required to obtain this information are working the same way.[130]

As far as bottom-up signals and top-down influences are concerned, incrementality could be readily integrated within the model. Chapter 5 presented a probabilistic

[130]It is important to make a distinction between incrementality and the dichotomy between pre-compositional and post-compositional derivation of implicatures. In Chapter 5, I briefly mentioned the differences between the two views, and we should note that incremental processing is compatible with these two views. See Foppolo & Marelli (2017); Breheny et al. (2013a,b) for more details about incremental processing of implicatures.

approach of top-down influences where the contextual-level representations were responsible for constraining the sentence-level conceptual space. The fact that this constraining process depends on the representational content at the contextual level implies that it would not be directly influenced by incrementality, i.e., the nature of the top-down influence would remain the same. However, in taking into account incrementality, we would have to update the contextual representations incrementally, and this is where incrementality could pose a problem.

As described in Chapter 5, the Rational Speech Act model usually considers whole utterances as inputs (Cohn-Gordon et al., 2019). To update the topic model and the situation model incrementally, we need to have a model of coordination that can tackle such increments at the contextual level. In line with what we just described for the 'syntactic game' (Skyrms, 2010), where a signal was divided into sub-signals, Cohn-Gordon et al. (2019) presented an incremental implementation of RSA where utterances are divided into sequences of linguistic units. I will not present this approach here because it would be beyond the scope of this thesis, but it is interesting to note that incremental approaches in pragmatics are at the heart of modeling linguistic prediction.

Another consequence of having an incremental model relates to the temporality associated with the linguistic prediction. When using an incremental model to update the contextual representations, we also need to acknowledge that at every moment t_i, the activation level of any words within the activation-based semantic network is a function of the activation level at time t_{i-1}. In Chapter 4, I explained that when a word is encountered, it activates other words related to it, and the connection weight between the two is derived from their similarity value. We can compare this process with a flow of activation that spreads throughout the semantic network (Rotaru et al., 2018).

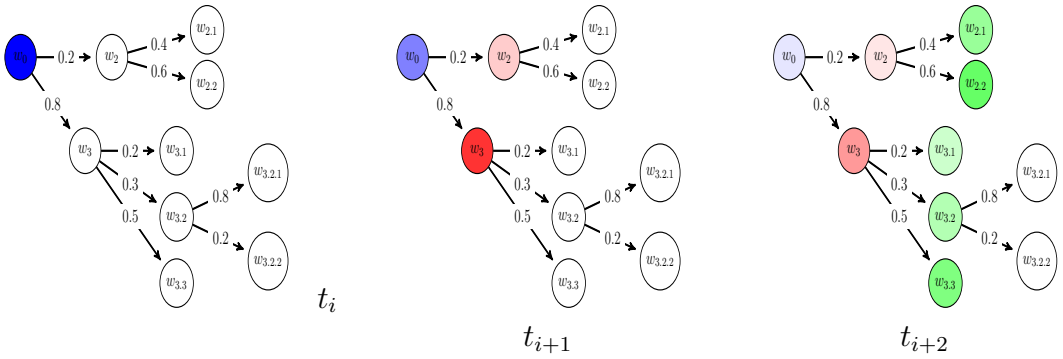


Figure 6.3 Depiction of the Spreading Activation at three consecutive moment t

In the non-incremental model presented here, the network reaches a global state where all the activations have distributed over time. In an incremental model, the spreading activation is incremented at every step t_i, and the flow of activation might not have had enough time to distribute throughout. To illustrate this incrementality in the flow of activations, we can use Figure 6.3 which is similar to the Figure representing the activation-based semantic network I presented in Chapter 4.

When a word w_0 is processed at time t_i, it activates w_1 and w_2 at time t_{i+1} with respect to their respective connection weights with w_0. At the time t_{i+2}, the activation spreads again and reaches the third level of activation. In an incremental model of linguistic prediction, the activation level of a word depends on the moment t. In contrast, in the non-incremental version of the model implemented in this thesis, the activation levels used to derive the prediction are the accumulation of the activations at all times. Furthermore, if we consider that the activation level of a word fades away over time, taking into account incrementality would increase the importance of information locality when it comes to a linguistic prediction, and this would be in line with the results from Futrell (2019).

The different treatment of the activation levels of the incremental and the non-

incremental model is also reminiscent of the distinction between the synchronous and the asynchronous binding, respectively, associated with the LISA and DORA model (Doumas et al., 2018; Doumas & Martin, 2018). Asynchronous binding allows representing a composition in terms of temporal firing patterns, much like what we just described for the incremental linguistic prediction model. Besides being highly similar to human cortical signals (Martin & Doumas, 2017), asynchronous binding seems to work well with an incremental treatment of activation-based semantic networks.

Finally, another aspect when considering incrementality is that we could have intra-sentential influences, e.g., the P-level representation might influence the activation at the RB-level and vice-versa. To model these kinds of influences, we can use the DORA-like time series that we presented in Chapter 4. In Figure 6.4 (Doumas et al., 2008, Figure 3), we see that the activation of the different sentential units is represented temporally. However, when modeling incremental linguistic prediction, the activation does not start at the P-level but at the lexical level because the first processed units are the words themselves. In other words, to model the incremental intra-sentential influences on the activation-based semantic network, we could use a time series centered around the lexical units.

6.1.3 Processing Cost

A third interesting factor that has to be considered is the processing cost involved in a linguistic prediction. It is essential to differentiate between the cost required to produce an input, which is usually associated with the speaker, and the pro-

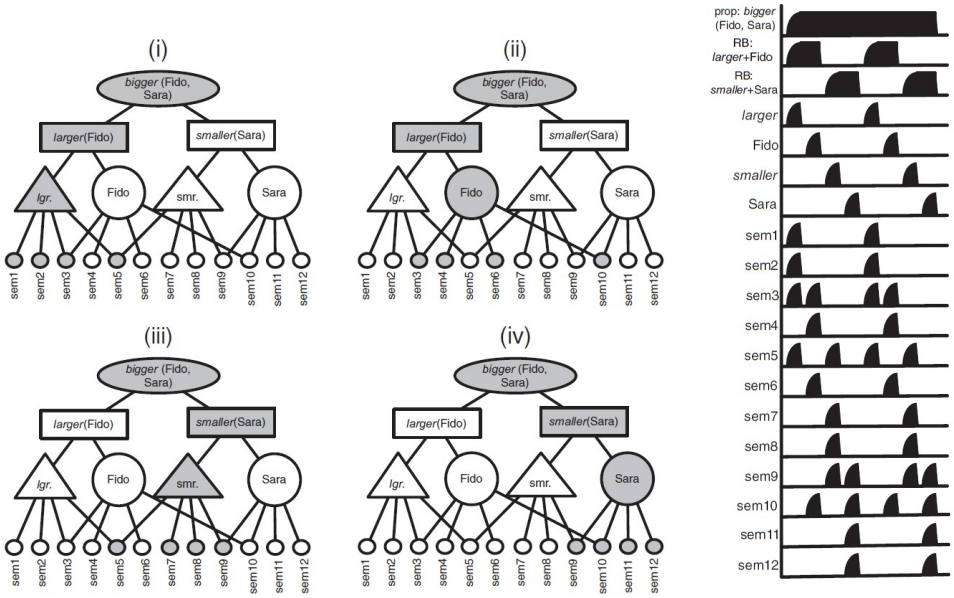


Figure 6.4 Left-side: Representation of the proposition *bigger(Fido,Sarah)* at four different time in the DORA model. Right-side: Time-series illustration of this representation

cessing cost of actually deriving a prediction, which would be associated with the interpreter.

The production cost of an utterance needs to be taken into account during the bottom-up derivation of the contextual representations. When the speaker chooses the best utterance to convey the intended meaning, production cost must be considered because the speaker wants to balance production cost and effectiveness. All things being equal, a speaker is more likely to choose a less costly utterance than an utterance with a very high production cost. When interpreting an utterance, the hearer uses this information about the production cost to infer the conveyed meaning from the speaker's perspective.

Brothers & Kuperberg (2021) differentiate between two kinds of processing cost on the interpreter's side: the first kind is related to the inherent difficulty of

processing a specific word and the second kind has to do with the difficulty of integrating the interpretation of this word into the contextual-level representations. This distinction has been supported by empirical evidence that associates each processing cost with a different neural response: an N400 effect when it comes to lexical access, and an increased late posterior positivity P600 effect for a violation involving higher-level representations (Kuperberg et al., 2020).

Concerning linguistic prediction, the first kind of processing cost is related to an upcoming word's unexpectedness correlated with its prior activation level. For example, when a word is not expected, this word is not activated within the activation-based semantic network, and this implies that it requires more cognitive effort to activate it from scratch when encountered. In comparison, when a word has already been activated within the interpreter's semantic network, the effort necessary to increase its activation level will be lower.

The second kind of processing cost has to do with updating the representations at the contextual level. It is a measure of the unexpectedness of the change that needs to occur at the contextual level to correspond with the latest incoming input that is processed. Since both the topic and the situation models are described as probabilistic distributions, the processing cost for integrating new information can be considered proportional to the discrepancy between the probability distributions before the update and after the update. Thus, a processing cost is the amount of cognitive energy required to change one's mental configurations.

Both of these processing costs are related to surprisal (Hale, 2001; Levy, 2008) because they are measures of the difficulty of processing a given piece of information, be it lexical or supra-lexical. These are measures of the cognitive effort required to integrate new information with respect to the hearer's expectation. In other words, this surprisal is a measure of the discrepancy between a prediction and an

input and not a measure of the cognitive cost required to generate a prediction (Delaney-Busch et al., 2019). However, if we consider the incremental aspect of linguistic prediction that we just described, the two become related because the surprisal at time t_i becomes the processing cost at time t_{i+1}.

If we recall the noisy-context surprisal model (Futrell et al., 2020; Futrell & Levy, 2017), the processing difficulty or the processing cost is proportional to the word's surprisal with respect to the noisy representation of the preceding context. In terms of our model of linguistic prediction, this means that when a new input is processed, the difficulty of integrating it within the representation of world knowledge at this time t is proportional to the degree by which it deviates from the probabilistic representation of the world knowledge at time t_{i-1}.

We can illustrate the processing cost's effect by looking at the three examples in (153) where we have two target positions per sentence. In terms of surprisal at the first target word, we can say that *fly* should be less surprising than *paint* since the latter is not usually associated with the context described by the first part of the sentence. Then, when the first target is *paint*, the surprisal at the second target position should be higher for the word *lamppost* compared with the word *picture* because it is much more common to paint a picture than to paint a lamppost. In terms of processing cost, (153-c) should thus be the sentence with the highest processing cost because it has two unexpected words.

(153) a. The day was breezy, so she went to the park to *fly* a *kite*.
b. The day was breezy, so she went to the park to *paint* a *picture*.
c. The day was breezy, so she went to the park to *paint* a *lamppost*.

Integrating these costs into the model of linguistic prediction presented in this

thesis requires the transposition of these unexpected changes at the various representational levels. At the first target position, the word *fly* should be easier to predict in the sense that its activation level should be higher than the one for *paint* since the topic model, the situation model, and the sentence-level representations would all contribute more to the activation of *fly* over the activation of *paint*.

Now, if we were to ask participants to complete (153-b) after *paint a*, the word *picture* would certainly be predicted more often than the word *lamppost*. This could be explained by the fact that the linguistic prediction at the second target position is more influenced by the preceding word *paint* than by the broader context of being in a park on a breezy day. In addition to the argument that the local context has more weight than the broader context, the linguistic prediction is also derived after the processing of an unexpected word, which means the cognitive effort to integrate the word *paint* at the contextual level has already been made. Consequently, it will have shifted those representations to fit this unexpected input, and it would then need an even bigger cognitive effort to shift it back to accommodate *lamppost*.

In terms of linguistic prediction, we can say that the hearer follows a path of least effort when predicting and that the processing cost is linked with the change to the probabilistic representations. In the case of the sentence-level representations, this processing cost is proportional to the difference in activation level. In the case of contextual representations, this change could be thought of as the difference between two probabilistic distributions.[131]

---

[131] In the last section of this chapter, I go further by linking the processing cost to an energy-based mechanism.

6.1.4 Individual differences

The third important issue that we ought to discuss is the individual differences between participants and how these differences could be modeled. Chapter 1 discussed the inter-individual variability in the cloze scores, and I mentioned that the cloze values are generally understood as average preferred responses. However, the fact that different individuals predict different words for the same sentence shows that the linguistic prediction might not be derived the same way by everybody.

This variability does not call into question the structural organization of the representations used in the linguistic prediction. Still, it illustrates that these representations might not be derived or used the same way by all individuals. Beyond the standard uncertainty involved in the derivation of these contextual level representations, each person would derive them using the same kind of bottom-up signal, and the effect of the higher-level representations would still work out as top-down influences. These inter-individual differences can be explained in terms of different knowledge about the world or in terms of different conceptual spaces. Between-population differences, like children and older adults, can be explained partly by this discrepancy of knowledge (Ryskin et al., 2020). This knowledge discrepancy can be accounted for in our linguistic prediction model by utilizing different word embeddings for different individuals or by using different mapping functions between these word embeddings and the contextual level representations.

Another option to explain this inter-individual variability is to consider the differential contribution between the contextual-level and the sentence-level representations. It might be possible that the contribution coming from top-down influences varies from individuals to individuals, and many empirical results seem to point

in that direction. Using an eye-tracking study, Huettig & Janse (2016) showed that individuals with higher working memory abilities and faster processing speed had more anticipatory eye movements. Similarly, Brothers et al. (2020) results are in line with the idea that skilled readers are better at detecting coherence breaks during online processing because they deal better with higher-level representations. Finally, Payne & Federmeier (2019) showed that individuals with a faster reading pace had larger predictability effects on the N400 components, which seems to be correlated to their capacity to take into account top-down contextual constraints; conversely, older adults (mean age of 68) are less likely to use top-down contextual constraints, and they rely more firmly on the sentence level representations (Payne & Federmeier, 2018).

Putting all this together, it seems that some individuals are better at taking into account higher-level representations. This ability to use top-down contextual constraints has been associated with a limitation of domain-general executive resources (Ryskin et al., 2020; Delaney-Busch et al., 2019). Even though I will not discuss these limitations in detail because it would be beyond the scope of this thesis, we can still mention how we could integrate such effects into this model of linguistic prediction. The simplest way to model this differential contribution from the sentence level and the contextual level is to define a Top-to-Bottom (TtB) ratio representing the relative ratio of the contribution between these two levels, as in (154).

(154) $$\text{TtB} = \frac{\text{Contribution from the contextual-level}}{\text{Contribution from the sentence-level}}$$

This Top-to-Bottom ratio is similar to a ratio of interpretation concerning the different kinds of interpretations presented in Brothers et al. (2020), namely a deep interpretation involving higher-level representations and a surface interpretation

does not involve higher representational levels. For example, if older adults rely more on the sentence level, then the top-down influence could be diminished by a corresponding factor, let us say 60%. Adding this TtB ratio would modulate the top-down constraining process at the source by modifying the probability distribution effect over the conceptual space.

Another crucial aspect of this Top-to-Bottom ratio is that it could also be relevant when comparing linguistic predictions performed in two different languages or two different cultures that speak the same language. For example, specific languages might be more under-determined than others, which might increase the importance of the contribution from the contextual level. Similarly, two populations might have different TtB ratios depending on their respective social and cultural characteristics of language use.

6.2 Related Approaches

To better contextualize the model of linguistic prediction presented in this thesis, it is crucial to discuss how it compares with other approaches. I present a connectionist approach that also uses a context-like representation. I also discuss possible connections between my model and the ever-growing number of predictive processing approaches, which view the brain as constantly generating and updating a multi-layered mental model of the representation of world knowledge.

6.2.1 Connectionnist Approaches

In a series of articles, Rabovsky et al. presented a computationally explicit account of the prediction process underlying the N400 amplitudes (Rabovsky, 2020; Rabovsky & McClelland, 2020; Rabovsky, 2019; Rabovsky et al., 2018, 2016). Instead of predicting words directly, they used Bayesian surprisal (Levy & Gayler, 2008) to produce N400 amplitudes so they could compare them with empirical results. Notably, a feature of their model is that they consider that prediction at the word level reflects changes happening at a higher level of representation. They allow predictions at the word level, but they state that the N400 component is triggered by representational updates in the integrated meaning of the sentence (Rabovsky et al., 2018; Rabovsky, 2019).

According to (Rabovsky et al., 2018), sentences are conveying information about situations or events, and a representation of a sentence should thus contain probabilistic information about the aspects of this situation or event (Rabovsky et al., 2018). In their model, they manipulate the representation of an event described by a sentence rather than representations of sentences directly. They call this representational level the "sentence-gestalt" (SG). The SG is a distributed representation containing all the information present in the sentence (St. John & McClelland, 1990), i.e., it represents the meaning a speaker wants to convey, and the form of this implicit SG-level representation is not constrained whatsoever.

In their model, they used distributed representations (DR) to represent concepts from a "pattern of activity over a collection of neurons" (Plate, 2006, p.2). In local representation, a concept is linked with a singular value or a scale. In a distributed representation, every concept is linked with the activation of more than one neuron, and conversely, every neuron can be linked with more than one concept

(Hinton et al., 1986). For simplicity reasons, I leave out the implementation side of things and only consider that a neuron is a state-like dimension that can be activated or not. A distributed representation has functional properties far more precise than non-distributed representations because it is easy to alter their properties slightly to emphasize certain features of a representation (St. John & McClelland, 1990), whereas it is not as straightforward to do so with isolated units. Also, distributed representations are ideal candidates for representing semantic meaning because they possess internal structure and carry more information than a single node (Chalmers, 1990). Finally, distributed representations allow for information to be processed simultaneously, and they are much closer to being able to represent a continuum (St. John & McClelland, 1990).

The fact that this representation is implicit allows representing aspects of meaning that are not explicitly expressed by the sentence. For example, if an event involves cutting a steak, then it would be understood that a knife is involved even if it is not explicitly expressed in the sentence. In other words, the meaning associated with the knife would be present in this implicit representation of the sentence even though the knife is not mentioned directly in the sentence. This is why we say that this implicit representation is a description of an event, and the SG-level represents the sentence describing this event. Instead of predicting a word directly, their model predicts which SG representation is the most plausible one and then predicts the word that would befit this most plausible distributed representation for the SG, much like what we did when comparing P-units. Their SG model supports the idea that thematic roles are fundamental constituents of events and that a representation of an event is constituted from pairs of thematic roles and words filling this role (St. John & McClelland, 1990).

> The Sentence Gestalt model holds that sentences constrain an implicit probabilistic representation of the meanings speakers to intend

to convey through these sentences. (Rabovsky et al., 2018, p.23).

In this connectionist model, we have to learn the SG-level representations to predict the upcoming word incrementally. These SG-level representations are derived from the meaning of the words expressed by the proposition, which, in turn, is represented as a sum of semantic features. This is very similar to what we had for the DORA model. Additionally, just like ours, these semantic features are based on McRae et al. (2005) empirically derived features (Rabovsky & McRae, 2014). For example, the meaning of the word *write* would be associated with the semantic features: is_an_action, is_done_with_letters, and is_productive.

To map the sentences' semantic features to the properties of the events, the connectionist approach uses a large number of sentences and events described in advanced (St. John & McClelland, 1990). If the number of matching pairs is large enough, then the system can discover the regularity of this matching, e.g., when "the man" is followed by a transitive verb in the active voice, the sentence refers to an event where "the man" is an agent.

The learning procedure involves learning the mapping between the events (semantic representation of the words) and the sentence representation. The idea is not to explicitly list all possible sentences and match them against all potential events, but to generate an appropriate response to given sentences so that the model can build a probabilistic distribution of the possible events linked with this sentence (Rabovsky et al., 2018). In other words, the goal of the learning phase is to generate sentence-event description pairs probabilistically. On one side, we have sentences composed of words, and on the other side, we have event descriptions that are a set of queries and associated responses. In this model, queries are related to the thematic role of the word concept, and it could, in general, be very large in scope and encompass other kinds of meaning (Rabovsky et al.,

2018). Also, Rabovsky et al. (2018) chose to model this prediction implicitly instead of explicitly because they did not want to constrain in any way the form of the representation even if the implicit representation is indirectly constrained by the choice of probes and queries. Instead of determining the exact structure (thematic roles and fillers) of an event, this implicit representation contains an ensemble of aspects represented as an ensemble of queries about the event, where each query is being associated with an ensemble of possible responses (Rabovsky et al., 2018). As depicted in Figure 6.5 (Rabovsky et al., 2018, Figure 1), the interaction between the input layer and the first hidden layer can be described as bottom-up constraints (also called 'constraints vectors') coming from each word to determine how each of these words influences the evolution of the SG-level (St. John & McClelland, 1990).

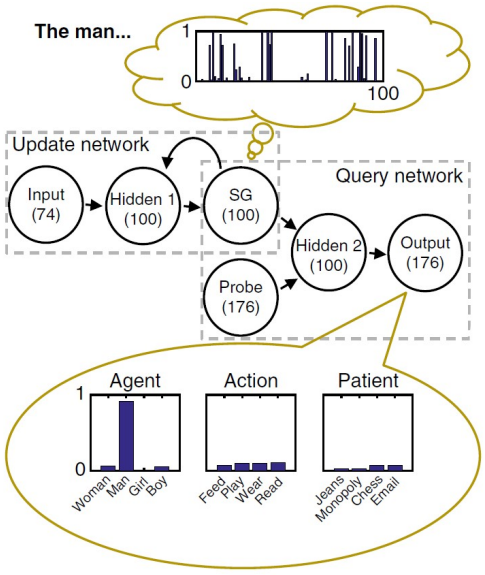


Figure 6.5 The sentence gestalt (SG) model architecture

The goal of the learning phase is to be able to match the probability distribution over possible answers to queries with this SG-level representation. Once this

phase has been performed, this method can recover probabilistic distributions for sentences it has never seen before. In such cases, the model tries to match this new sentence to an event using the same regularities it has observed and incorporated during the training. From there, we can use the retrieved SG representation to model changes incrementally. When the model encounters a word n, then the implicit representation is updated accordingly by minimizing the divergence between the prior distribution and the posterior distribution that fits the encountered word best. Their model learned the implicit representations using a close world containing a finite number of possible words.[132]

Rabovsky's model bears some resemblance to what I have described for the DORA and the LISA hierarchical structure, even though the latter is a hybrid symbolic-connectionist approach. However, in Rabovsky's model, the training phase that determines the dynamics of activation given the input words is purely connectionist, making it more challenging to make sense of a single event's compositional features.

An essential difference between their model and the model that I described here is that they trained their model from a 'simple microworld' (Rabovsky et al., 2018, p.693). That is to say that all words that are used during the training are fixed in the model and, most importantly, the probabilities for every event, i.e., the priors, are also artificially pre-determined.[133] Training representations from a microworld give an excellent opportunity to model the effect of contextual information that is usually difficult to graphs formally. This is a bit different from the approach I presented here because I used the similarity space derived from

---

[132]See Rabovsky et al. (2018) for the complete list of words and events.

[133]For their model of the influence of the world knowledge on linguistic processing, Venhuizen et al. (2019) followed a similar path when training their representations using a predetermined microworld.

word embeddings models without any exterior influence related to the plausibility of a particular event. Using this similarity space as a starting point allows us to construct a similarity matrix built directly from a corpus of data without pre-fixing the influence coming from the context.

Finally, in Rabovsky's model, the probability that a particular word is chosen as a continuation is computed from the SG-level representation only. In my model, I am considering different contributions in parallel: the semantic features, the lexical units, the RB-units, and the P-units. Despite these few differences, Rabovsky's model is an up-and-coming model that uses SG representations to translate between sentences and events, and they described events as consisting of action, a location, and some thematic roles like agent and patient. It does not presuppose strong compositionality by letting the interactions between SG-level and word-level determine how the SG-level is updated. Another advantage of this approach is that it allows top-down influences to override bottom-up activation from a word whenever the contextual constraints are very strong St. John & McClelland (1990) which is to say that the contribution from the word differs depending on the representation at the SG-level (McClelland et al., 1989), much like what I have described in this thesis.

6.2.2 Predictive Processing

I now discuss the Predictive Processing view and its relationship with what I presented in this thesis.[134] When performing a cloze task, participants are asked to

[134]Predictive processing is sometimes described as Prediction Error Minimization or PEM, or even more strictly as predictive coding (Hohwy, 2018). Here, the term *predictive processing* encompasses approaches for which prediction acts to minimize error when processing informa-

predict the upcoming word using the information that is accessible to them at the point of truncation. This predictive process can be categorized as optional because it is not compulsory by nature to predict the missing word. However, according to the Predictive Processing approach, predicting is automatic and unavoidable.

Predictive features of the brain have been highlighted in neuroscience (Friston, 2008; Kveraga et al., 2007; Kunde et al., 2007; Bubic et al., 2010), and, in recent years, it has been recognized that prediction is a much more ubiquitous process than simply completing a cloze task. Predictive features encompass terms like prediction, anticipation, and expectation, and they refers to the idea that the brain is naturally inclined to predict the following upcoming input. Predictive processing helps minimize cognitive effort and speeds up computation (Clark, 2013; Hohwy, 2013; Clark, 2016), i.e., if the input matches the prediction, then the amount of cognitive effort needed to process the input is reduced Friston et al. (2016).

These approaches generally assume that predictive processing is also valid at the implementation level (Marr, 1982) and that it is thus possible to regroup cognition, perception, action, and attention under a common framework (Clark, 2013; Bubic et al., 2010). Basically, following the predictive processing hypothesis, the brain does not passively wait for the following input, but it actively keeps adjusting to the kind of information it expects to process next Kveraga et al. (2007); Friston et al. (2016). In other words, it is the predictive processes that drive the actions to be realized, and, conversely, any actions are necessarily triggered by an anticipated response Kunde et al. (2007).

This idea of predicting in advance to minimize the cognitive effort spent when processing new information is in line with the good-enough and now-or-never pro-

---

tion.

cessing views I discussed in Chapter 5 because the goal of anticipation is to prepare the mind or the brain to process new information (Ferreira & Chantavarin, 2018). Given the limited amount of cognitive resources, this preparation facilitates the integration of this new information (Ferreira & Lowder, 2016; Luke & Christianson, 2016). Additionally, predicting the upcoming information in advance allows it to be readily processed, i.e., to begin the chunk-and-pass operation as early as possible (Christiansen & Chater, 2016).

Generally speaking, a prediction at a lower representational level is derived by using higher-level representations. In this thesis, I described the inter-level relationship between the sentence level and the contextual level, but there could be many more levels involved, and each of these levels could be treating its kind of information at its own specific time-scale (Hohwy, 2016). Within this multi-level environment, the information flows from one level to the other using top-down generative models and bottom-up inputs (Clark, 2013). These top-down generative signals derive a prediction at a lower level which is then compared with the bottom-up input.

If the prediction is correct, the input is processed most rapidly because the system is already prepared. On the other hand, in cases where the prediction is not accurate, an error signal is sent to update the representations to fit the newly encountered information Kuperberg & Jaeger (2016). To generate such predictions, the brain uses the specific context of the situation to activate associations of related representations by finding analogies between the input and memory Kunde et al. (2007). The main difficulty with this approach is that to predict the upcoming information, a hearer must know which world model is correct. In other words, the brain has to predict a piece of information that it has not encountered yet (Hohwy, 2018). One way to understand how this could work is to use the free-energy principle, which is used in statistical physics and theoretical biology,

to model the self-preservation of the energy of an organism (Friston, 2009, 2008; Friston et al., 2016, 2017). This principle guides the conservation of the neural energy available in the brain (Friston, 2010).

Friston et al. (2017) described the brain as a self-organizing inference machine that actively predicts and is dedicated to optimizing predictions. These predictions are generated from top-down signals that are then compared with bottom-up sensory inputs, and the goal of the brain is to optimize the predictions, so the discrepancy between them and the upcoming input is minimized (Friston, 2010). These predictions could be derived using Bayesian inferences that combine prior probability distributions with observations to compute a prediction about the posterior probability distribution (Jones & Love, 2011; Aitchison & Lengyel, 2017; Zeevat, 2015a).

$$P(w|s) = \frac{P(w)P(s|w)}{P(s)} \tag{155}$$

This probability of w given a state s is generated and then compared with the following upcoming input. According to Friston (2010), to minimize the surprisal at one level, one has to adjust one's own beliefs at a higher representational level. In other words, this difference between the representations at a particular level is responsible for the updates at the levels above it. This process is called "active inference" (Friston et al., 2016), and it corresponds to the inference that arises as to the consequence of minimizing the free-energy (Friston et al., 2015). In other words, the active inference updates the representations levels to minimize the free-energy, which is, in itself, equivalent to minimizing surprisal (Mirza et al., 2016). This approach based on the free-energy principle is fascinating because it models the brain as a physical dynamical system. Although I will not describe Friston's model in its entirety because it would be beyond this thesis's scope,

it is interesting to note that it aligns with what I described in Chapter 5 when discussing the effect of top-down influences on conceptual spaces.

This predictive processing mechanism guides perception and actions, but as stated by Clark (2013), language differs from both action and perception, which means that the transposition to language processing might not be that straightforward. Even though Friston et al.'s active inference model might not have been primarily designed for modeling language processing, more recent implementations are now beginning to integrate linguistically related simulations Friston et al. (2017).[135]

When it comes to language, the top-down prediction is about pre-activating the lexical features of upcoming inputs with respect to their expected likelihood (Brothers & Kuperberg, 2021) much like what we discussed throughout this thesis. The hierarchical model described by Friston seems to fit well with the structure of the model of linguistic prediction described in this thesis because it already has a hierarchical structure where a transition at a higher level entails a sequence of transitions at lower levels (Friston et al., 2017).

However, in the multi-level representational model we presented in Chapter 5, the bottom-up signals contribute to the derivation of the higher-level representations, and the top-down influences are constraining the processes happening at lower levels. The main goal of this model is to predict the following upcoming information. In comparison, under the predictive processing perspective, the top-down signals are compared with the bottom-up inputs being processed. Error signals are then generated from the discrepancy between these two opposing signals. The system aims to optimize the representational levels to minimize these prediction error signals, i.e., to minimize the surprisal. Within this perspective, every level is

---

[135]See Friston et al. (2017) for a presentation of a mixed model aimed at simulating reading.

hierarchically accountable (Friston, 2010). If the representation at a given level, say the contextual level, is updated for some reason, e.g., from new contextual information, then the linguistic processing at lower levels will be constrained accordingly.

Another difference between the two is that the multi-layer model presented here is not strictly hierarchical because the four compositional levels are treated independently when it comes to their contribution to the contextual level representations. In contrast, the usual depiction of a predictive processing architecture is strictly linearly hierarchical. Every representational level sits between two other representational levels, and the error signal is generated from the interaction of a top-down signal coming from a higher level and a bottom-up input coming from below. Notwithstanding this structural difference, the linguistic prediction structure I presented in this thesis has been developed to be compatible with the predictive processing architecture. In this thesis, I remained agnostic regarding the hierarchical organization of the different intra-level representations involved at the sentence and contextual levels. However, it could be possible to rearrange them to fulfill the stricter hierarchical requirements of the predictive processing account.

The simple parallel I trace here between linguistic prediction and predictive processing opens the door for broader integration of these predictive processes within more general models of linguistic processing. Concerning pragmatic processing, considering predictive processing could enable us to revisit the nature of some pragmatic phenomena in terms of predictive processes. Additional pragmatic factor could include, the implicature, the amount of prior context, , and presupposition, to name a few. The growing interest in these predictive processing approaches will undoubtedly influence the development of different research fields interested in linguistic information processing.

6.3 Looking Forward: A Topological Model of Linguistic Prediction

In this thesis, I used Bayesian inferences to model both bottom-up signals and top-down influences, but other approaches involving the geometry of the conceptual space are also very promising. The purpose of this last section is to revisit our model of linguistic representations in terms of a topological approach to conceptual space.[136] This discussion is exploratory, and it is by no means an exhaustive introduction to topological and energy-based approaches.

6.3.1 Top-down Influences

Top-down influences are a natural candidate to be modeled using a topological approach because they are directly responsible for modifying the configuration of the conceptual space. Contextual influences affect the conceptual space in a non-monotonic way, and the changes to the conceptual space reverberate as modifications to the categorization of different concepts within this conceptual space. This idea is in line with the results from Roth & Shoben (1983) where they got different typicality judgments when immersed in different contexts. Similarly, Barsalou & Sewell (1984) showed that categorization was influenced by the participant's point of view, i.e., by their global context. When an individual changes its own categorization, the conceptual space regions are modified: the quality dimensions attached to these regions are either changing or weighted differently (Khater & Tawfik, 2009).

---

[136] Topology is a mathematical approach interested in the preservation of the properties of an object that undergoes a continuous deformation like bending, stretching, twisting, etc.

Gärdenfors & Williams (2001) presents three different ways to model the contextual modifications applied to a conceptual space. The basic premise is to posit that the context modifies the conceptual space itself. This modification can be done by adding or removing certain concepts or prototypes or by changing the boundaries and the categorical regions (Gärdenfors, 2000).[137]

One way to modify the conceptual space is to directly change the distance between regions/concepts, affecting the boundaries between regions (Gärdenfors & Williams, 2001). The problem with this avenue is that it does not seem to fit the contextual representations I described. For example, when discussing the topic model's effect, I mentioned that the topic influences the probability distribution for the possible continuations, but this influence does not modify the set of possible continuations, only their relative activation. Another problem with this approach is related to the fact that both the situation model and the topic model do not change our general conceptual space, but they adapt it to the context. In other words, during a linguistic prediction, we do not forget or create concepts, i.e., we keep on using what we already know, but we select our conceptual knowledge that is the most relevant to the situation at hand.

A second approach is to change the mapping between objects and regions of the conceptual space (Gärdenfors & Williams, 2001). Here, instead of modifying the weights of the different regions, the context would change the concepts' position within the conceptual space. Gärdenfors & Williams (2001) illustrate this idea by discussing the different mappings for a word like *Tweety*. If we know that *Tweety* has feathers and two wings and is very close to the generic prototype of a bird, then it would be mapped onto the same region as other generic birds,

---

[137]These three approaches also work when considering semantic networks and the activation-based account we discussed in Chapter 4 because we could apply it to the similarity space from which the activation-based network is derived.

but if we were to learn new information about it, e.g., that it cannot fly, then we would re-map it onto the non-flying bird region containing the emu and the kiwi for example. In a case like this, when the global context is updated with new information about the meaning of a word, the mapping between this word and the conceptual space is also updated. Here, updating the context thus modifies the mapping of a word while keeping the regions of the conceptual space unchanged.

A third way to think of the context's influence on the conceptual space is to keep the same conceptual space and narrow down the number of accessible regions. In other words, when immersed in a context, we would only be able to access a small region of the whole conceptual space. This operation could be performed by combining or bounding some regions. It could also be done by adding new weights to the quality dimensions so that the regions remained the same even though the similarity between them would be modified. Using this approach would have the effect of contracting some regions while dilating others (Gärdenfors & Williams, 2001), thus changing the distance between two points in the conceptual space (Khater & Tawfik, 2009) without modifying the content of the conceptual space. This contraction-dilatation operation on the conceptual space is illustrated in Figure 6.6 (Gärdenfors & Williams, 2001, Figure 5). In Figure 6.6, the left side represents the conceptual space at a prior time, and the right side represents it at a posterior time. The dotted line represents the prior configuration of the conceptual space.

As we can see in Figure 6.6, the content of the conceptual space remains the same, i.e., we still have p_1, p_2, p_3, and q, and it is the relationship between these points, and their relationship to the space that changes.[138]

---

[138]For those interested in the mathematical properties of the model, if we consider that the conceptual space is a manifold, we could see these changes as changes about the metric or the geometry of this manifold.

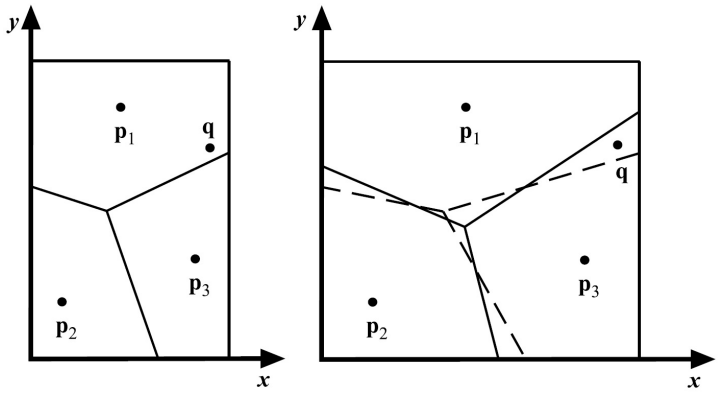


Figure 6.6 Dilation and Contraction of the Conceptual Space

This third approach is very interesting because it conceptualizes contextual effects as geometrical constraints imposed on the conceptual space. Furthermore, modifying the mapping between conceptual space and a concept could well be modeled as a change in the weights to the quality dimensions because the resulting conceptual space would be the same. However, the converse is not true because changing a mapping is not like narrowing down to a smaller region of the conceptual space, even though the resulting state after changing a mapping could be equivalent to a space where the salience weights between the regions have been altered.

Going back to our topic model, we could say that the contextual representations modify the probability distribution for possible continuations because a given topic increases the salience of the words related to it by bringing them closer in the conceptual space. In other words, a topic acts like an attractor that brings words closer together if they are related to it.

This method based on the geometrical aspect of the conceptual space, as elegant as it can be, is very complex to implement, but one way to model these top-down influences could be to use *Attractor Networks* (Hopfield, 1982) which are a set of nodes sitting on a D-dimension space that allows mapping a continuous

space to a discrete space. In attractor networks, the nodes are activated when there is enough appropriate input to activate them.[139] Even if these networks are biologically plausible (Baggio & Hagoort, 2011), they are difficult to work with, and they are computationally demanding to implement (Zemel et al., 2008).[140]

6.3.2 Conceptual Spaces and Energy

Once we obtain a bent conceptual space, we can then use a dynamic energy-based approach to model linguistic prediction and other kinds of transitions involving different states of the world. Energy-based topological models can be viewed as a generalization of probabilistic models (Lecun et al., 2019) which implies we can represent probabilistic states of the world as local minima separated by potential energy barriers as in Figure 6.7. Energy-based models associate scalar energy to each configuration of the variables (Lecun et al., 2019), and, under this perspective, a prediction is derived by finding the best local minimum corresponding to the input at a given time t.

In Figure 6.7 the configuration of the variables correspond to the representation of world knowledge s_i, and the scalar energy is related to the probability of a given state within a particular context, i.e., the representation of world knowledge that is the most preferable, or the most relevant. We can understand this as a correspondence between the contextual representations and the expected repre-

---

[139]Some modern implementations of these networks can be added as a layer during deep-learning to provide new ways to train models (Ramsauer et al., 2020).

[140]Another promising topological approach is related with sheaving construction where meaning is grounded by the topological space (Phillips, 2020). See Phillips (2018); Abramsky & Brandenburger (2011) for more information about this approach.

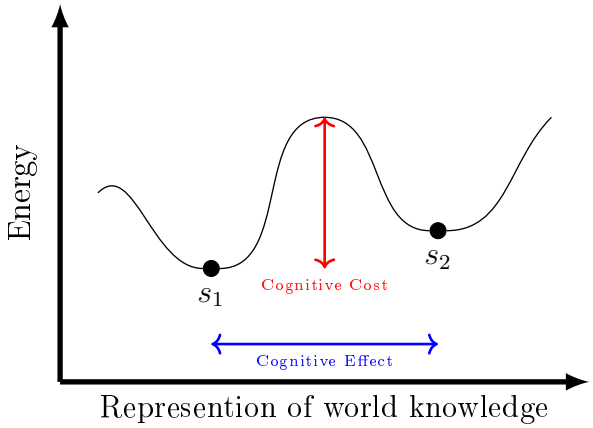


Figure 6.7 Probabilistic States of the World Represented as Local Minima

sentation of world knowledge, where a predictive inference consists of finding the representation of world knowledge that minimizes the scalar energy (Lecun et al., 2019).

For cases when the prediction is inaccurate, the hearer is forced to transit from a local minimum to another local minimum by following a path of minimum effort. This means that when the representation of world knowledge is updated to fit this new input, the interpreter chooses the easiest cognitive path, i.e., transiting through the smallest energy barrier available. If the gain of information is worth it compared to the processing cost, the update is triggered. On the contrary, if the gain of information is too small, the cognitive effort is not worth it, and the representation of world knowledge does not change. As shown in Figure 6.7, the cognitive cost is the height of the potential barrier between two contiguous states of the world, and the cognitive effect is the difference in the position of the two statess_1 to s_2, which is determined from the content of the these states. For example, if s_2 contains more information than s_1, the cognitive effect will be

positive if we go from s_1 to s_2.[141]

Even though there are still several aspects of this topological and energy-based approach to work out, the capacity to use energy or the space metric to conceptualize the influence of the context and the transition between the different states of the worlds during linguistic processing is very promising. The use of local attractors or local minima could help explain non-commutativity linguistic issues where the derivation of a particular interpretation is dependent on the order in which the information is presented to the hearer. For example, modeling linguistic interpretation as a dynamical process happening in a topological conceptual space might help better understand the role of information locality during linguistic processing since the transition between states of the world would happen locally because global transitions would require much more cognitive effort. To sum up, an energy-based approach would not change the nature of the representational structure, but it would certainly change the way these signals are modeled.

[141] Considerations about this processing cost for transiting between states are naturally integrated within Friston et al. (2015) predictive processing model in the form of an expected cost derived from an expected gain of information.

CONCLUSION

This thesis develops a new theoretically-driven model that predicts the anticipation of upcoming words in a sentence. Using a highly interdisciplinary perspective regarding the nature of linguistic prediction and the kinds of cognitive processes involved therein, I present a set of cognitive desiderata that linguistic theories must consider: incrementality, non-monotonicity, and interpretability of sub-propositional content. I differentiate two kinds of contributions when deriving a linguistic prediction: those coming from different levels of semantic granularity and those coming from the coordination of linguistic interaction, and I present a language model that marries these two contributions. Finally, empirical adequacy was assessed by the three worked-out examples for which the theory match the ordering that was obtained empirically.

Throughout the thesis, I discussed different kinds of truncated sentences. For example, in (156-a), the presence of the word *loaf* is a strong indicator that the missing word is *bread* because the two have very high statistics of co-occurrence. On the other hand, in (156-b), the information expressed in the first part of the sentence leaves much room in terms of possible continuations because there are many things a kind old man can ask us to do.

(156) a. I went to the bakery for a loaf of ...

b. The kind old man asked us to ...

To better understand how this linguistic prediction is performed, I chose to divide it in terms of the questions we have to answer:

1. How can we determine the realm of possible continuations from the information available to the participant?

2. How can we represent the information that is available to the participant when deriving a prediction?

The first question is related to the correspondence function that allows the interpreter to derive the lexical items that best fit the truncated sentence. In the approach described in this thesis, the realm of possible continuations is obtained using an activation-based semantic network where the level of activation of any concepts at a particular time represents the degree by which they are triggered by the information retrieved from the truncated sentence and the global context.

This relative value of the spreading activation is proportional to the connection weight between these concepts, which can be treated like a probability of co-occurrence between these two words. At any given time, these co-occurrence probabilities determine the linguistic prediction based on the relationships between all the concepts represented in the semantic network. The mapping function that goes from information, i.e., a state of the world, to a linguistic prediction, i.e., a lexical item, in this case, is thus derived from these co-occurrence activation-based probabilities.

These relationships between the words and their associated connection weights can be obtained from different methods. In this thesis, I derived the semantic network from similarity matrices representing the similarity of co-occurrence between different linguistic constructions. We thus have one similarity matrix for the lexical items, one for the RB-units, one for the P-units, and one for the SF-units. These similarity matrices are computed using a distributional approach representing words as n-dimension vectors that can be directly compared using mathematical operations such as the cosine similarity measure. One great ad-

vantage of representing word-meaning as vectors is that many different learning algorithms can extract these vectors from massive corpora. On the other hand, the main downside is that all the downstream computation depends on these word embeddings and how they are obtained. In this thesis, I used word2vec word embeddings as a basis for the derivation of the similarity matrices. However, I could have used other training algorithms as well.[142]

Once we have a way to assign a relative probability of occurrence for potential continuations, we can turn to the second question related to retrieving the information available to the hearer right before the linguistic prediction. As I already described, this available information is equivalent to the state of the world the interpreter conceives herself to be in at that moment. However, to derive the representation of this state of the world, we must consider both the contribution from the truncated sentence and the contribution from the global context.

Chapter 3 distinguished two aspects of linguistic interaction, the compositional aspect, and the coordination aspect. The coordination aspect is related to the fact that when a speaker wants to convey meaning using a given utterance, this speaker must consider the hearer's perspective to maximize the chance that the utterance is correctly interpreted. Besides, the interpreter must also consider the perspective of the speaker when retrieving the expressed meaning. On the other hand, the compositional aspect of linguistic communication is related to the expression of complex meaning as different linguistic units. This compositional aspect is crucial because it is by combining words and compounds that complex meaning emerges.

The compositional aspect of linguistic meaning involves both syntax and seman-

---

[142]Word embedding models are constantly improving their capability to extract linguistic information from language use, and the model presented here is not committed to a particular kind of word embeddings.

tics. In Chapter 2, I presented empirical evidence that supported the idea that syntax and semantics were two separate processing streams and that the semantic stream usually has precedence over the syntactic stream. Only when the meaning of a word is almost absent, i.e., when we encounter a new unknown word, the syntactic stream's contribution might supersede the one from the semantic stream. Keeping this in mind, in Chapter 4, I decided to focus on the contribution from the semantic aspect of the composition, which is not to say that syntax does not contribute to the derivation of the linguistic prediction. The contribution from the syntax does not play the same role as does the semantic meaning of a composition, but it still has a role to play because it is constraining at the level of the grammatical type of the upcoming word. In contrast, the semantic stream plays a role at the semantic type level, i.e., what kind of meaning should be expected.

I use this semantic information as a basis for my model of linguistic prediction. Furthermore, the semantic information available to the hearer when a composition is processed is decomposed into four types of representations.

(157) The cup is bigger than the ball.

a. P-unit: bigger(cup,ball)

b. RB-units: larger+cup, smaller+ball

c. Lexical-level: cup, bigger, ball

d. Semantic features: size, attribute, high, low, dish, mug, kitchen, ...

After having processed (157), the representation of world knowledge in the mind of the hearer, i.e., the activation level of the concepts represented in the semantic network, consists of the superposition of the information presented at every level of the sentence. The meaning of a composition is the combination of all the meanings expressed by the sentence-level units. In the case of (157), the compositional

meaning triggers the linguistic units expressed at these four levels. In other words, when we hear that the cup is bigger than the ball, not only are we representing this as a P-unit, but we also process semantic information at the sub-propositional level like that the cup is larger and the ball is smaller, and that that size is a semantic feature of this composition.

In this thesis, I presented a language model where a linguistic prediction is derived from the combination of the contributions from these four levels and where each level triggers an activation signal that spreads throughout the semantic network. This language model is strictly about a linguistic prediction in a cloze task setting, which means this model's output is the predictability value discussed in Chapter 2. As discussed in Chapter 4, the problem concerning the concepts of similarity, predictability, and plausibility being all interrelated and difficult to separate is thus evacuated in my language model because the predictability obtained from this language model encompasses both the similarity and the plausibility values. In this model, the linguistic prediction is derived from the similarity values for different sentence-level representations, and plausibility is transposed into a similarity value between two propositions or between two P-units.

Chapter 3 described how to model the coordination aspect of linguistic interaction using the existing tools offered by different approaches to pragmatics, but we still had to find a way to represent the contextual information that affects the derivation of a linguistic prediction. For this thesis, I created two kinds of models to represent contextual information: the topic and situation models. I presented a multi-layered representation of linguistic prediction that integrates the contribution from the sentence-level representations, the contribution from the contextual level, and the constant interaction between them.

From the bottom-up, the sentence-level representations are responsible for the

derivation of these contextual models because the only information available to the hearer is expressed in the truncated sentence. Therefore, we have to transpose linguistic information into contextual information using Bayesian inferences where the higher-level representations were derived probabilistically from the lower-level ones. The topic model is derived from a pre-trained topic distribution space representing the relationship between topics and words. The situation model is derived from the speaker's perspective using these sentence-level representations as building blocks.

These contextual representations then influence the predictive process by constraining the linguistic prediction. Consequently, continuations that are not supported by the context are inhibited within the semantic network. This means that we can keep the same derivation process as before because it is not the prediction itself that is influenced by the contextual representations but the conceptual space upon which the activation levels of potential prediction are derived. Given the interaction between these levels, we see that both have a primordial role in the derivation of any prediction. This structural model is centered around the coordination aspect of linguistic interaction, and it illustrates the crucial connection between the representational levels involved in pragmatic processing.

Chapter 2 discussed the phonological, syntactic, and semantic streams' independence, but the pragmatic stream is often either wholly omitted or integrated within the semantic stream. In this thesis's model, the pragmatic stream influences the syntactic and semantic streams, as depicted in Chapter 5. Moreover, the contribution from this pragmatic stream or its effect on linguistic prediction is primarily due to the structure of processing responsible for treating this information. The particularity of this pragmatic stream is about the interrelated representational levels. Especially the fact that to transfer the information from one level to another, we ought to consider the other agent involved and our knowledge about the

world.

The syntactic stream and the semantic stream are context-free processes regarding the information treated as an input and the one given as an output, i.e., the nature of the process does not change with respect to the context. In other words, no matter the global context we are in, the syntactic stream transforms a word-token into a syntactical type, and the semantic stream transforms a word-token into a semantic type as in the processing architecture from Baggio (2018).

This principle of stability regarding linguistic processing should also be valid for the pragmatic stream. Generally speaking, the most exciting thing about the context is not the correspondence between a particular context and a sentence but the general structure of this correspondence. The model developed in this thesis is in line with the idea that we could, in principle, determine a processing structure that could be applied to any situation, i.e., context-free contextual processing. No matter the context we are in, both the bottom-up signals and the top-down influences are involved. This means that the focus on pragmatic processing should not be about the relationship between a particular input and its output but about the general processes by which this input is transformed into this output. Conceptualizing the pragmatic stream as a processing structure is one step towards developing a context-free theory of pragmatics.

BIBLIOGRAPHY

Abramsky, S. & Brandenburger, A. (2011). The sheaf-theoretic structure of non-locality and contextuality. *New Journal of Physics*, *13*

Aitchison, L. & Lengyel, M. (2017). With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227

Allott, N. & Textor, M. (2012). Lexical Pragmatic Adjustment and the Nature of Ad hoc Concepts. *International Review of Pragmatics*, *4*(2), 185–208

Allott, N. N. (2006). Game Theory and Communication. In A. Benz, G. Jäger, & R. van Rooij (eds.), *Game Theory and Pragmatics* pp. 123–151. London: Palgrave Macmillan UK

Ambati, B. R., Deoskar, T., Johnson, M. & Steedman, M. (2015). An incremental algorithm for transition-based CCG parsing. *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pp. 53–63.

Ambati, B. R., Reddy, S. & Steedman, M. (2016). Assessing relative sentence complexity using an incremental CCG parser. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pp. 1051–1057.

Armeni, K., Willems, R. M. & Frank, S. L. (2017). Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience and Biobehavioral Reviews*, *83*, 579–588

Aurnhammer, C. & Frank, S. L. (2019a). Comparing Gated and Simple Recurrent Neural Network Architectures as Models of Human Sentence Processing. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pp. 112–118.

Aurnhammer, C. & Frank, S. L. (2019b). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*(September), 107198

Bach, K. (1997). The Semantics-Pragmatics Distinction: What it is and Why it Matters. *Linguistische Berichte*, *8*, 33–50.

Bach, K. (1999). The myth of conventional implicature. *Linguistics and Philosophy*, *22*, 262–283.

Bach, K. (2001). Semantically speaking. In I. Kenesei & R. M. Harnish (eds.), *Perspectives on Semantics, Pragmatics, and Discourse*. Amsterdam: John Benjamins.

Bach, K. (2004). Minding the Gap. In C. Bianchi (ed.), *The Semantics/Pragmatics Distinction* pp. 27–43. CSLI Publications.

Bach, K. (2012). Saying, meaning, and implicating. In K. Allan & K. M. Jaszczolt (eds.), *The Cambridge Handbook of Pragmatics* pp. 47–68. Cambridge: Cambridge University Press

Baggio, G. (2018). *Meaning in the Brain*. The MIT Press

Baggio, G. & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, *26*(9), 1338–1367

Baggio, G., Stenning, K. & Lambalgen, M. V. (2019). Semantics and cognition. In M. Aloni & P. Dekker (eds.), *The Cambridge Handbook of Formal Semantics* pp. 756–774. Cambridge: Cambridge University Press

Baggio, G., van Lambalgen, M. & Hagoort, P. (2008). Computing and recomputing discourse models: An ERP study. *Journal of Memory and Language*, *52*(1), 36–53

Balkir, E., Kartsaklis, D. & Sadrzadeh, M. (2015). Sentence Entailment in Compositional Distributional Semantics. *arXiv:1512.04419 [cs, math]*

Barker, C. & Shan, C.-C. (2014). *Continuations and Natural Language*. Oxford University Press.

Baron, S. G. & Osherson, D. (2011). Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage*, *55*(4), 1847–1852

Baroni, M. (2013). Composition in distributional semantics. *Linguistics and Language Compass*, *7*(10), 511–522

Baroni, M., Bernardi, R. & Zamparelli, R. (2014a). Frege in Space: A Program of Compositional Distributional Semantics. *Linguistic Issues in Language Technology*, *9*(6), 242–346.

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language resources and evaluation*, *43*(3), 209–226.

Baroni, M., Dinuand, G. & Kruszewski, G. (2014b). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings for the 52nd Annual Meeting for the Association of Computational Linguists*, pp. 238–247.

Barrett, J. A. (2009). The evolution of coding in signaling games. *Theory and Decision*, *67*(2), 223–237

Barsalou, L. W. & Sewell, D. R. (1984). *Constructing representations of categories from different points of view.* Technical report, Emory Cognition Project Technical Report #2, Emory University.

Bayes, T. (1764). An Essay Toward Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Benz, A., Jäger, G. & van Rooij, R. (2006). *Game Theory and Pragmatics.* Palgrave Macmillan.

Benz, A., Jäger, G., van Rooij, R., Rooij, R. V., van Rooij, R., Rooij, R. V. & van Rooij, R. (2005). An introduction to game theory for linguists. In G. J. &. v. R. e. A. Benz (ed.), *Game Theory and Pragmatics* pp. 293. Houndsmills, Basing-stoke, Hampshire: Palgrave Macmillan

Bergen, L., Goodman, N. D. & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society*, pp. 120–125.

Bergen, L., Levy, R. & Goodman, N. D. (2014). Pragmatic Reasoning through Semantic Inference. *Semantics & Pragmatics*, *9*(1984), 1–43

Berkum, J. J. (2013). Anticipating communication. *Theoretical Linguistics*, *39*(1-2), 75–86

Binmore, K. (2009). *Rational Decisions.* Princeton University Press

Blei, D. M., Griffiths, T. L. & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, *57*(2), 1–30

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Block, C. K. & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, *42*(3), 665–670

Bloom, P. A. & Fischler, I. (1980). Completion norms for 329 sentence contexts. *Memory & cognition*, *8*(6), 631–642

Blutner, R. (1998). Lexical pragmatics. *Journal of Semantics*, *15*(2), 115–162

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*

Bornkessel-Schlesewsky, I. & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews*, *59*(1), 55–73

Bornkessel-Schlesewsky, I. & Schlesewsky, M. (2019). Toward a Neurobiologically Plausible Model of Language-Related, Negative Event-Related Potentials. *Frontiers in Psychology*, *10*(February), 1–17

Bornkessel-Schlesewsky, I., Schlesewsky, M., Small, S. L. & Rauschecker, J. P. (2015). Neurobiological roots of language in primate audition: common computational properties. *Trends in Cognitive Sciences*, *19*(3), 142–150

Bos, J. (2015). Open-Domain Semantic Parsing with Boxer. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pp. 301–304., Vilnius, Lithuania. Linkoping University Electronic Press, Sweden.

Boudewyn, M. A., Long, D. L. & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension. *Cognitive, Affective Behavioral Neuroscience*, *15*(3), 607–624

Bransford, J. D., Barclay, J. & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology*, *3*(2), 193–209

Breheny, R., Ferguson, H. J. & Katsos, N. (2013a). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes*, *28*(4), 443–467

Breheny, R., Ferguson, H. J. & Katsos, N. (2013b). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition*, *126*(3), 423–440

Brehm, L., Jackson, C. N. & Miller, K. L. (2019). Speaker-specific processing of anomalous utterances. *Quarterly Journal of Experimental Psychology*, *72*(4), 764–778

Brennan, J. R. & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, *14*(1), e0207741

Brothers, T. & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*(September 2020), 104174

Brothers, T., Swaab, T. Y. & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language*, *93*, 203–216

Brothers, T., Wlotko, E. W., Warnke, L. & Kuperberg, G. R. (2020). Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language*, *21*(1), 1–58

Brouwer, H., Crocker, M. W., Venhuizen, N. J. & Hoeks, J. C. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*, *41*, 1318–1352

Brouwer, S., Mitterer, H. & Huettig, F. (2012). Discourse context and the recognition of reduced and canonical spoken words. *Applied Psycholinguistics*, *34*, 1–21

Bubic, A., von Cramon, D. Y. & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*(March), 1–15

Bunt, H. (2016). Computational Pragmatics. In Y. Huang (ed.), *The Oxford Handbook of Pragmatics*, volume 1. Oxford University Press

Burnard, L. (2007). Reference Guide for the British National Corpus (XML Edition). Retrieved from `http://www.natcorp.ox.ac.uk/XMLedition/URG/`

Calmus, R., Wilson, B., Kikuchi, Y. & Petkov, C. I. (2020). Structured sequence processing and combinatorial binding: neurobiologically and computationally informed hypotheses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *375*(1791)

Carston, R. (2002). *Thoughts and Utterances.* Malden, MA, USA: Blackwell Publishing

Carston, R. (2004). Relevance Theory and the Saying/Implicating Distinction. In L. Horn & G. Ward (eds.), *The Handbook of Pragmatics*, volume 1054 pp. 633–658. Oxford, UK: Oxford Blackwell

Carston, R. (2012). Relevance Theory. In *Routledge Companion to Philosophy of Language* pp. 1–20. Routledge

Carston, R. & Powell, G. (2006). Relevance Theory - New Directions and Developments. In E. Lepore & B. C. Smith (eds.), *Oxford Handbook of Philosophy of Language* pp. 341–360. Oxford University Press.

Casasanto, D. & Lupyan, G. (2015). All Concepts Are Ad Hoc Concepts. In E. Margolis & S. Laurence (eds.), *The Conceptual Mind: New directions in the study of concepts* pp. 543–566. Cambridge (Mass.): MIT Press

Chalmers, D. (1990). Why Fodor and Pylyshyn were wrong: The simplest refutation. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 340–347.

Chalmers, D. J. (1993). Connectionism and compositionality: Why Fodor and Pylyshyn were wrong. *Philosophical Psychology*, *6*(3), 305–319

Champollion, L., Alsop, A. & Grosu, I. (2019). Free choice disjunction as a rational speech act. *Semantics and Linguistic Theory*, *29*, 238

Chater, N. & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344

Chater, N., Tenenbaum, J. B. & Yuille, A. (2006a). Probabilistic models of cognition: Conceptual foundations. *Trends Cogn Sci*, *10*(7), 287–291

Chater, N., Tenenbaum, J. B. & Yuille, A. (2006b). Probabilistic models of cognition: where next? *Trends in Cognitive Sciences*, *10*(7), 292–293

Chernov, G. V. (2004). *Inference and Anticipation in Simultaneous Interpreting: A probability-prediction Model.* Johh Benjamins.

Chow, W. Y., Momma, S., Smith, C., Lau, E. & Phillips, C. (2016a). Prediction as memory retrieval: timing and mechanisms. *Language, Cognition and Neuroscience*, *31*(5), 617–627

Chow, W. Y., Smith, C., Lau, E. & Phillips, C. (2016b). A "bag-of-arguments" mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, *31*(5), 577–596

Christiansen, M. H. & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action and the Embodied Mind*. Oxford, UK: Oxford University Press.

Clark, H. H. (1996). *Using Language*. Cambridge, MA: Cambridge University Press

Clark, R. (2012). *Meaningful games: Exploring Language with Game Theory*. MIT Press.

Clark, S. (2015). Vector Space Models of Lexical Meaning. In S. Lappin & C. Fox (eds.), *The Handbook of Contemporary Semantic Theory* pp. 493–522. Chichester, UK: John Wiley & Sons, Ltd

Clark, S., Coecke, B. & Sadrzadeh, M. (2008). A compositional distributional model of meaning. In *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, pp. 133–140. College Publications

Clarke, D. (2012). Challenges for Distributional Compositional Semantics. *arXiv preprint arXiv:1207.2265*

Coecke, B., Sadrzadeh, M. & Clark, S. (2010). Mathematical Foundations for a Compositional Distributional Model of Meaning. *Linguistic Analysis*, *36*, 345–384

Cohn-Gordon, R., Goodman, N. D. & Potts, C. (2019). An Incremental Iterated Response Model of Pragmatics. *Proceedings of the Society for Computation in Linguistics*, *2*

Collins, A. M. & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428

Connolly, J. F. & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience*, *6*(3), 256–266

Corbeil, M. (2014). *Relevance Theory and Communication: On the differences between speaker and hearer (MRes Dissertation)*. University College London, London.

Corbeil, M. (2015). Dynamique de la cognition et interprétation du sens. In *Actes du XXIX colloque Les Journées de la Linguistique*.

Corps, R. E., Gambi, C. & Pickering, M. J. (2018). Coordinating Utterances During Turn-Taking: The Role of Prediction, Response Preparation, and Articulation. *Discourse Processes*, *55*(2), 230–240

Corps, R. E., Pickering, M. J. & Gambi, C. (2019). Predicting turn-ends in discourse context. *Language, Cognition and Neuroscience*, *34*(5), 615–627

Cosentino, E., Baggio, G., Kontinen, J. & Werning, M. (2017). The time-course of sentence meaning composition. N400 effects of the interaction between context-induced and lexically stored affordances. *Frontiers in Psychology*, *8*(MAY), 1–17

Culicover, P. W. & Jackendoff, R. (2006). The simpler syntax hypothesis. *Trends in Cognitive Sciences*, *10*(9), 413–418

Curran, J. R. (2003). *From Distributional to Semantic Similarity Doctor of Philosophy*. University of Edinburgh.

Dawkins, R. (2006). *The Selfish Gene*. Oxford University Press

De Jaegher, K. (2008). The evolution of Horn's rule. *Journal of Economic Methodology*, *15*(3), 275–284

De Ruiter, J. P., Mitterer, H. & Enfield, N. J. (2006). Projecting the end of a Speaker's Turn: A Cognitive Cornerstone of Conversation. *Source: Language*, *82*(3), 515–535

Delaney-Busch, N., Morgan, E., Lau, E. & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition*, *187*(January), 10–20

DeLong, K. A., hsuan Chan, W. & Kutas, M. (2019). Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology*, *56*(4), 1–14

DeLong, K. A. & Kutas, M. (2020). Comprehending surprising sentences: sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. *Language, Cognition and Neuroscience*, *3798*

DeLong, K. A., Quante, L. & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*(1), 150–162

Delong, K. A., Troyer, M. & Kutas, M. (2014). Pre-Processing in Sentence Comprehension: Sensitivity to Likely Upcoming Meaning and Structure. *Linguistics and Language Compass*, *8*(12), 631–645

DeLong, K. a., Urbach, T. P. & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, *8*(8), 1117–1121

Demberg, V. & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*

Ding, N., Melloni, L., Tian, X. & Poeppel, D. (2017). Rule-based and word-level statistics-based processing of language: insights from neuroscience. *Language, Cognition and Neuroscience*, *32*(5), 570–575

Doumas, L. A. & Hummel, J. E. (2012). Computational Models of Higher Cognition. *The Oxford Handbook of Thinking and Reasoning*

Doumas, L. A., Hummel, J. E. & Sandhofer, C. M. (2008). A Theory of the Discovery and Predication of Relational Concepts. *Psychological Review*, *115*(1), 1–43

Doumas, L. A. & Martin, A. E. (2018). *Learning structured representations from experience* (1 éd.), volume 69. Elsevier Inc.

Doumas, L. A., Morrison, R. G. & Richland, L. E. (2018). Individual differences in relational learning and analogical reasoning: A computational model of longitudinal change. *Frontiers in Psychology*, *9*(JUL), 1–14

Doumas, L. A. A., Hamer, A., Puebla, G. & Martin, A. E. (2017). A theory of magnitude and similarity detection and learning. In *39th annual conference of the Cognitive Science Society (CogSci 2017)*, volume 1, pp. 1955–1960.

Douven, I. & Gärdenfors, P. (2020). What are natural concepts? A design perspective. *Mind and Language*, *35*(3), 313–334

Dowty, D. (2007). Compositionality as an Empirical Problem. In C. Barker & P. Jacobson (eds.), *Papers from the Brown University Conference on Direct Compositionality* pp. 23–101. Oxford University Press

Engel, A. K., Fries, P. & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top–down processing. *Nature Reviews Neuroscience*, *2*(10), 704–716

Eppley, B. L. (2014). Dependency-Based Word Embeddings. *Soft-Tissue Surgery of the Craniofacial Region*, pp. 351–358.

Erkan, G. & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479

Federmeier, K. D. & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, *41*(4), 469–495

Fedorenko, E., Blank, I. A., Siegelman, M. & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *203*(November 2018), 104348

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database.

Ferreira, F. & Chantavarin, S. (2018). Integration and Prediction in Language Processing: A Synthesis of Old and New. *Current Directions in Psychological Science*, *27*(6), 443–448

Ferreira, F. & Lowder, M. W. (2016). *Prediction, Information Structure, and Good-Enough Language Processing*, volume 65. Elsevier Ltd

Ferreira, F. & Swets, B. (2002). How Incremental Is Language Production? Evidence from the Production of Utterances Requiring the Computation of Arithmetic Sums. *Journal of Memory and Language*, *46*(1), 57–84

Firth, J. R. (1957). A Synopsis of Linguistic Theory. *Studies in Linguistic Analysis*, pp. 1–32.

Fitzsimmons, G. & Drieghe, D. (2013). How fast can predictability influence word skipping during reading? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1054–1063

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge, MA.: MIT Press.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71

Fodor, J. D. & Ferreira, F. (eds.) (1998). *Reanalysis in Sentence Processing*, volume 21 of *Studies in Theoretical Psycholinguistics*. Dordrecht: Springer Netherlands

Foppolo, F. & Marelli, M. (2017). No delay for some inferences. *Journal of Semantics*, *34*(4), 659–681

Frank, M. C. & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998

Frank, S. L., Haselager, W. F. & van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition*, *110*(3), 358–379

Frank, S. L., Otten, L. J., Galli, G. & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11

Frank, S. L. & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203

Franke, M. (2009). *Signal to Act: Game Theory in Pragmatics*. University of Amsterdam.

Franke, M. (2010). Semantic meaning and pragmatic inference in non-cooperative conversation. *Interfaces: Explorations in Logic, Language and Computation*, pp. 13–24.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, *4*(1), 1–82

Franke, M. (2013). Game Theoretic Pragmatics. *Philosophy Compass*, *8*(3), 269–284.

Franke, M. (2016). The Evolution of Compositionality in Signaling Games. *Journal of Logic, Language and Information*, *25*(3-4), 355–377

Franke, M. & Bergen, L. (2020). Theory-driven statistical modeling for semantics and pragmatics: A case study on grammatically generated implicature readings. *Language*, *96*(2), e77–e96

Franke, M., De Jager, T., Rooij, R. V., van Rooij, R. & Rooij, R. V. (2012). Relevance in Cooperation and Conflict. *Journal of Logic and Computation*, *22*(1), 23–54

Franke, M. & Degen, J. (2015). Reasoning in Reference Games: individual vs. population-level probabilistic modeling. *PLoS ONE*, *11*(5), 1–49

Franke, M. & Jäger, G. (2013). Pragmatic Back-and-Forth Reasoning. *Pragmatics, Semantics and the Case of Scalar Implicatures*, pp. 170–200.

Franke, M. & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Special Issue: Formal pragmatics. Zeitschrift für Sprachwissenschaft*, *35*(1), 1–30

Frege, G. (1884). *The Foundations of Arithmetic (Die Grundlagen der Arithmetik)*. Oxford: Blackwell.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, *100*, 25–50.

Frege, G. (1984). *Collected Papers on Mathematics, Logic and Philosophy*. Oxford: Blackwell.

Friston, K. (2005). A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci*, *1456*(360), 815–836

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, *4*(11)

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews. Neuroscience*, *11*(2), 127–138

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J. & Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews*, *68*, 862–879

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T. & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T. & Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Frontiers in Human Neuroscience*, *7*(September), 1–18

Friston, K. J., Parr, T. & de Vries, B. (2017). The graphical brain: Belief propagation and active inference. *Network Neuroscience*, *1*(4), 381–414

Futrell, R. (2019). Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pp. 2–15., Stroudsburg, PA, USA. Association for Computational Linguistics

Futrell, R., Gibson, E. & Levy, R. P. (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, *44*(3)

Futrell, R. & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, *1*(Section 2), 688–698

Futrell, R., Qian, P., Gibson, E., Fedorenko, E. & Blank, I. (2019). Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pp. 3–13., Stroudsburg, PA, USA. Association for Computational Linguistics

Gärdenfors, P. (1996). Philosophy and Cognitive Science: Categories, Consciousness, and Reasoning. *Philosophy and Cognitive Science: Categories, Consciousness, and Reasoning*, pp. 0–16.

Gärdenfors, P. (2000). *Conceptual Spaces: On the Geometry of Thought*. MIT Press.

Gärdenfors, P. (2004). Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, *2*(2), 9–27.

Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.

Gärdenfors, P. & Williams, M. A. (2001). Reasoning about categories in conceptual spaces. *IJCAI International Joint Conference on Artificial Intelligence*, pp. 385–392.

Gastaldi, J. L. (2020). Why Can Computers Understand Natural Language? *Philosophy & Technology*

Geurts, B. (2010). *Quantity Implicatures*. Cambridge: Cambridge University Press

Gibson, E., Bergen, L. & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(20), 8051–8056

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L. & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, *23*(5), 389–407

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. MIT Press.

Goodkind, A. & Bicknell, K. (2018). 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)*, pp. 10–18., Stroudsburg, PA, USA. Association for Computational Linguistics

Goodman, N. & Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, *5*(1), 173–184

Goodman, N. D., Baker, C. L. & Tenenbaum, J. B. (2009). Cause and Intent : Social Reasoning in Causal Learning. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*, pp. 2759–2764.

Goodman, N. D. & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829

Green, M. S. (1995). Quantity, volubility, and some varieties of discourse. *Linguistics and Philosophy*, *18*(1), 83–112

Grefenstette, E. (2013). *Category-Theoretic Quantitative Compositional Distributional Models of Natural Language Semantics*. University of Oxford.

Grefenstette, E., Sadrzadeh, M., Clark, S., Coecke, B. & Pulman, S. (2014). Concrete Sentence Spaces for Compositional Distributional Models of Meaning. In H. Bunt, J. Bos, & S. Pulman (eds.), *Computing Meaning. Text, Speech and Language Technology* pp. 71–86. Dordrecht: Springer

Grice, P. H. (1957). Meaning. *The philosophical Review*, *66*(3), 377–388.

Grice, P. H. (1989). *Studies in the Way of Words*. Cambridge, Mass.: Harvard University Press.

Griffiths, T. L., Kemp, C. & Tenenbaum, J. B. (2008). Bayesian Models of Cognition. In R. Sun (ed.), *The Cambridge Handbook of Computational Psychology* pp. 59–100. Cambridge University Press

Griffiths, T. L., Steyvers, M. & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244

Grisoni, L., Miller, T. M. C. & Pulvermüller, F. (2017). Neural correlates of semantic prediction and resolution in sentence processing. *Journal of Neuroscience*, *37*(18), 4848–4858

Hagoort, P. (2020). The meaning-making mechanism(s) behind the eyes and between the ears. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *375*(1791), 20190301

Hagoort, P., Brown, C. & Groothusen, J. (1993). The syntactic positive shift (sps) as an erp measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439–483

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*, pp. 1–8., Morristown, NJ, USA. Association for Computational Linguistics

Hale, J. (2016). Information-theoretical Complexity Metrics. *Language and Linguistics Compass*, *10*(9), 397–412

Halliday, M. A. (1967). Notes on transitivity and theme in English, part II. *Journal of Linguistics*, *3*(2), 199–244.

Harris, Z. (1954). Distributional Structure. *Word*, *10*(23), 146–162.

Hasson, U., Egidi, G., Marelli, M. & Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, *180*(June), 135–157

Heim, I. (1982). The semantics of definite and indefinite NPs. *University of Massachusetts at Amherst dissertation*, pp. 263.

Heim, I. (2008). File Change Semantics and the Familiarity Theory of Definiteness. In P. Portner & B. H. Partee (eds.), *Formal Semantics* pp. 223–248. Oxford, UK: Blackwell Publishers Ltd

Heller, D., Parisien, C. & Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition*, *149*, 104–120

Heylighen, F. (2005). *Towards an anticipation control theory of mind*. Technical report, Vrije Universiteit Brussel

Hinton, G. E., McClelland, J. & Rumelhart, D. (1986). Distributed representations. In D. Rumelhart & J. McClelland (ed.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations* pp. 77–109. Cambridge, MA.: MIT Press.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428

Hohwy, J. (2013). *The Predictive Mind*. Oxford, UK: Oxford University Press.

Hohwy, J. (2016). The Self-Evidencing Brain.Noûs, 2016. 50(2):. *Noûs*, *50(2)*, 259–285.

Hohwy, J. (2018). Prediction error minimization in the brain. In M. Sprevak & M. Colombo (eds.), *Routledge Handbook to the Computational Mind*. Oxford: Routledge.

Holler, J., Kendrick, K. H., Casillas, M. & Levinson, S. C. (2015). Editorial: Turn-Taking in Human Communicative Interaction. *Frontiers in Psychology*, *6*(December), 1–4

Holyoak, K. J. & Morrison, R. G. (2005). Thinking and Reasoning: A Reader's Guide. In K. J. Holyoak & R. G. Morrison (eds.), *The Cambridge Handbook of Thinking and Reasoning*. Cambridge University Press

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 79, pp. 2554–2558.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (ed.), *Meaning, Form, and Use in Context: Linguistics Applications*. Washington: Georgetown University Press.

Horn, L. (1995). Presupposition and implicature. In S. Lappin (ed.), *The Handbook of Contemporary Semantic Theory* pp. 299–319. Oxford: Blackwell.

Horn, L. (2010). Issues in the Investigation of Implicature. In *Meaning and Analysis: New essays on Grice* pp. 310–340. Palgrave.

Huang, Y. (2016). Neo-Gricean Pragmatics. In Y. Huang (ed.), *The Oxford Handbook of Pragmatics*, volume 1. Oxford University Press

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135

Huettig, F. & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience*, *31*(1), 80–93

Huettig, F. & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, *31*(1), 19–31

Hummel, J. E. (2011). Getting Symbols out of a Neural Architecture Achieving Independence and Role-Filler. *Connection Science*, *23*, 109–118

Hummel, J. E. & Holyoak, K. J. (2003). A Symbolic-Connectionist Theory of Relational Inference and Generalization. *Psychological Review*, *110*(2), 220–264

Hummel, J. E. & Holyoak, K. J. (2005). Relational Reasoning in a Neurally Plausible Cognitive Architecture. *Current Directions in Psychological Science*, *14*(3), 153–157

Ito, A., Corley, M. & Pickering, M. J. (2018). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism*, *21*(2), 251–264

Ito, A., Corley, M., Pickering, M. J., Martin, A. E. & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157–171

Ito, A., Gambi, C., Pickering, M. J., Fuellenbach, K. & Husband, E. M. (2020). Prediction of phonological and gender information: An event-related potential study in Italian. *Neuropsychologia*, *136*(November 2019), 107291

Jackendoff, R. (2007). A Parallel Architecture perspective on language processing. *Brain Research*, *1146*(1), 2–22

Jacobs, R. A. & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 8–21

Jaeger, G. (2008). Applications of Game Theory in Linguistics. *Language and Linguistics Compass*, *2*(3), 406–421.

Jaeger, T. F. & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57–83

Jäger, G. (2014). Rationalizable Signaling. *Erkenntnis*, *79*(S4), 673–706

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, Mass.: Harvard University Press.

Jones, M. & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169–188

Jurafsky, D. (2002). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod & J. Hay (eds.), *Probabilistic linguistics*. MIT Press

Kaiser, E. (2011). On the Relation between Coherence Relations and Anaphoric Demonstratives in German. *Proceedings of the 2010 Annual Conference of the Gesellschaft für Semantik. Sinn und Bedeutung 15*, pp. 337–351.

Kaiser, E. (2013). Looking beyond personal pronouns and beyond English: Typological and computational complexity in reference resolution. *Theoretical Linguistics*, *39*(1-2), 109–122

Kaiser, E. & Cherqaoui, B. (2016). Effects of coherence on anaphor resolution and vice versa : Evidence from French personal pronouns and anaphoric demonstratives. In A. Holler & K. Suckow (eds.), *Empirical Perspectives on Anaphora Resolution* pp. 51–78. Berlin, Boston: de Gruyter.

Kalina, S. (1992). Discourse processing and interpreting strategies: an approach to the teaching of interpreting. In C. D. Loddegaard & A. (eds.), *Teaching Translation and Interpreting. Training, Talent and Experience.* pp. 251–257. Amsterdam: Johh Benjamins.

Kamide, Y., Altmann, G. T. & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156

Kamp, H. (1981). A theory of truth and semantic representation. In J. Groenendijk & M. Stokhof (eds.), *Formal methods in the Study of Language* pp. 277–322. Mathematical Centre Tracts 135.

Kamp, H. (1995). Discourse representation theory. In *Handbook of Pragmatics* pp. 253–257. Amsterdam: John Benjamins Publishing Company

Kamp, H. (2008). A Theory of Truth and Semantic Representation. In *Formal Semantics* pp. 189–222. Oxford, UK: Blackwell Publishers Ltd

Kamp, H., Van Genabith, J. & Reyle, U. (2011). Discourse Representation Theory. In *Handbook of Philosophical Logic* pp. 125–394. Dordrecht: Springer Netherlands

Kao, J. T., Bergen, L. & Goodman, N. D. (2014a). Formalizing the Pragmatics of Metaphor Understanding. *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014)*, *1*, 719–724.

Kao, J. T. & Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. *Proceedings of the 36th Conference of the Cognitive Science Society*, pp. 1051–1056.

Kao, J. T., Wu, J. Y., Bergen, L. & Goodman, N. D. (2014b). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007

Karimi, H., Brothers, T. & Ferreira, F. (2019). Phonological versus semantic prediction in focus and repair constructions: No evidence for differential predictions. *Cognitive Psychology*, *112*(April), 25–47

Karuza, E., Thompson-Schill, S. L. & Bassett, D. (2017). Network traversal mediates expectations. *arXiv*

Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. CSLI Publications: Stanford University.

Kehler, A., Kertz, L., Rohde, H. & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, *25*(1 SPEC. ISS.), 1–44

Kehler, A. & Rohde, H. (2018). Prominence and coherence in a Bayesian theory of pronoun interpretation. *Journal of Pragmatics*

Kempson, R., Meyer-Viol, W. & Gabbay, D. M. (2001). *Dynamic Syntax: The Flow of Language Understanding*. Wiley-Blackwell.

Khachatryan, E., Brouwer, H., Staljanssens, W., Carrette, E., Meurs, A., Boon, P., Van Roost, D. & Van Hulle, M. M. (2018). A new insight into sentence comprehension: The impact of word associations in sentence processing as shown by invasive EEG recording. *Neuropsychologia*, *108*(December 2017), 103–116

Khater, M. F. & Tawfik, A. Y. (2009). Context evolution in conceptual spaces. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, pp. 147–150.

Kherwa, P. & Bansal, P. (2018). Topic Modeling: A Comprehensive Review. *ICST Transactions on Scalable Information Systems*, *7*(24), 159623

Kiela, D. & Clark, S. (2014). A Systematic Study of Semantic Vector Space Model Parameters. In S. Gothenburg (ed.). *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality*, volume 353, pp. 21–30., Gothenburg, Sweden. Association for Computational Linguistics

Kim, A. E., Oines, L. & Miyake, A. (2018). Individual differences in verbal working memory underlie a tradeoff between semantic and structural processing difficulty during language comprehension: An ERP investigation. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(3), 406–420

Kim, A. E., Oines, L. D. & Sikos, L. (2015). Prediction during sentence comprehension is more than a sum of lexical associations: the role of event knowledge. *Language, Cognition and Neuroscience*, *3798*(November), 1–5

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182

Koleva, N., Horbach, A., Palmer, A. & Ostermann, S. (2014). Paraphrase Detection for Short Answer Scoring. In *NEALT Proceedings Series*, volume 22, pp. 59–73., Uppsala, Sweden. LiU Electronic Press.

Körding, K. P. & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3-4), 243–276

Kunde, W., Elsner, K. & Kiesel, A. (2007). No anticipation-no action: The role of anticipation in action and perception. *Cognitive Processing*, *8*(2), 71–78

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*(1), 23–49

Kuperberg, G. R. (2013). The Proactive Comprehender: What Event-Related Potentials tell us about the dynamics of reading comprehension. In B. Miller, L. Cutting, & P. McCardle (eds.), *Unraveling the Behavioral, Neurobiological, and Genetic Components of Reading Comprehension*. Baltimore: Paul Brookes Publishing.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, *31*(5), 602–616

Kuperberg, G. R., Brothers, T. & Wlotko, E. W. (2020). A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35

Kuperberg, G. R. & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*.

Kutas, M., DeLong, K. A. & Smith, N. J. (2011). A Look around at What Lies Ahead: Prediction and Predictability in Language Processing. In *Predictions in the Brain* pp. 190–207. Oxford University Press

Kutas, M. & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647

Kutas, M. & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity

Kveraga, K., Ghuman, A. S. & Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain cognition*, *65*(2), 145–168

Lambek, J. (2008). From word to sentence. *Polimetrica, Milan*

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem : The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*(2), 211–240

Landauer, T. K., Folt, P. W. & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2), 259–284

Lau, E. F., Holcomb, P. J. & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience*, *25*(3), 484–502

Le Ny, J.-F. (1978). Psychosemantics and simultaneous interpretation. In D. G. Sinaiko & H.W. (eds.), *Language interpretation and communication* pp. 289–298. Plenum Press.

Leckey, M. & Federmeier, K. D. (2020). The P3b and P600(s): Positive contributions to language comprehension. *Psychophysiology*, *57*(7), 1–15

Lecun, Y., Chopra, S., Hadsell, R., Ranzato, M. A. & Huang, F. J. (2019). Energy-Based Models. *Predicting Structured Data*, pp. 1–59.

Lederer, M. (1978). Simultaneous interpretation: Units of meaning and other features. In D. G. &. H. W. Sinaiko (ed.), *Language interpretation and communication* pp. 323–333. New York: Plenum Press.

Lenci, A. (2008). Distributional approaches in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(May), 1–31.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, *4*(1), 1065975714

Levelt, W. J. M. (2012). *Speaking: From Intention to Articulation*, volume 66. MIT Press.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

Levinson, S. C. (2000). *Presumptive Meaning*. MIT Press / Bradford Books.

Levinson, S. C. (2016). Turn-taking in Human Communication; Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, *20*(1), 6–14

Levinson, S. C. & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6*(June), 1–17

Levy, O. & Goldberg, Y. (2014). Dependency-Based Word Embeddings. *ACL*, pp. 302–308.

Levy, O., Goldberg, Y. & Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: Formal techniques and empirical results. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *1*, 1055–1065.

Levy, S. D. & Gayler, R. W. (2008). Vector symbolic architectures: A new building material for artificial general intelligence. *Frontiers in Artificial Intelligence and Applications*, *171*(1), 414–418.

Lewis, D. (1969). *Convention*. Cambridge: Harvard University Press.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of philosophical logic*, *8*(1), 339–359

Li, Y., Bandar, B., McLean, D. & O'Shea, J. (2004). A method for measuring sentence similarity and its application to conversational agents. In *The 17th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*, pp. 820–825., Florida, USA. American Association for Artificial Intelligence (AAAI) Press.

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D. & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, *18*(8), 1138–1150

Lieto, A., Chella, A. & Frixione, M. (2016). Conceptual Spaces for Cognitive Architectures : A lingua franca for different levels of representation. *Biologically Inspired Cognitive Architectures*, pp. 1–9.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. *Proceedings of ICML*, pp. 296–304.

Linzen, T. & Jaeger, T. F. (2016). Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. *Cognitive science*, *40*(6), 1382–1411

Linzen, T. A. (2019). What can linguistics and deep learning contribute to each other? Response to pater. *Language*, *95*(1), e99–e108

Liu, Y. & Zong, C. (2004). Example-Based Chinese-English MT. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, pp. 6093–6096.

Luke, S. G. & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, *28*(2), 203–208

Lupyan, G. & Lewis, M. (2017). From words-as-mappings to words-as-cues: the role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *0*(0), 1–19

Mandera, P., Keuleers, E. & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78

Marcus, G. (2018). Deep Learning: A Critical Appraisal. *arXiv preprint arXiv: 1801.00631*, pp. 1–27.

Marr, D. (1982). *Vision*. San Francisco, Ca: W. H. Freeman and Company.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*(5417), 522–523

Marslen-Wilson, W. & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1–71

Martin, A. E. (2019). A compositional architecture for language built on neural oscillations. *Psyarxiv*.

Martin, A. E. & Baggio, G. (2020). Modelling meaning composition from formalism to mechanism. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190298

Martin, A. E. & Doumas, L. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLoS Biology*, *15*(3), 1–23

Martin, A. E. & Doumas, L. A. (2019). Predicate learning in neural systems: using oscillations to discover latent structure. *Current Opinion in Behavioral Sciences*, *29*, 77–83

Martin, A. E. & Doumas, L. A. (2020). Tensors and compositionality in neural systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791)

McClelland, J., Rumelhart, D. & Hinton, G. (1986). The Appeal of Parallel Distributed Processing. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, pp. 3–44.

McClelland, J. L., St. John, M. & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, *4*(3-4), SI287–SI335

McCready, E. (2014). *Reliability in Pragmatics*. Oxford University Press

McRae, K., Cree, G. S., Seidenberg, M. S. & Mcnorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559

Michalon, O. & Baggio, G. (2019). Meaning-driven syntactic predictions in a parallel processing architecture: Theory and algorithmic modeling of ERP effects. *Neuropsychologia*, *131*(May), 171–183

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arxiv:1301.3781*, pp. 1–12.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C. & Joulin, A. (2017). Advances in Pre-Training Distributed Word Representations. *arXiv preprint arXiv:1712.09405*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546*, pp. 1–9.

Mikolov, T., Yih, W. T. & Zweig, G. (2013c). Linguistic regularities in continuous spaceword representations. *NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference*, pp. 746–751.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, *38*(11), 39–41.

Milward, D. (1995). Incremental interpretation of Categorial Grammar. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics* -, volume 53, pp. 119., Morristown, NJ, USA. Association for Computational Linguistics

Mirza, M. B., Adams, R. A., Mathys, C. D. & Friston, K. J. (2016). Scene Construction, Visual Foraging, and Active Inference. *Frontiers in Computational Neuroscience*, *10*(June)

Mitchell, J. & Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, *34*(8), 1388–1429

Moeschler, J. (2013). Is a speaker-based pragmatics possible? Or how can a hearer infer a speaker's commitment? *Journal of Pragmatics*, *48*(1), 84–97

Moore, G. E. (1953). *Some Main Problems of Philosophy*. Routledge

Morgan, E. U., van der Meer, A., Vulchanova, M., Blasi, D. E. & Baggio, G. (2020). Meaning before grammar: A review of ERP experiments on the neurodevelopmental origins of semantic processing. *Psychonomic Bulletin & Review*

Morrill, G. (2000). Incremental processing and acceptability. *Computational Linguistics*, *26*(3), 318–338

Nicenboim, B., Vasishth, S. & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia*, *142*(April)

Nieuwland, M. S. (2019). Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience & Biobehavioral Reviews*, *96*(October 2018), 367–400

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Husband, E. M., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., Rueschemeyer, S.-A. A., Segaert, K., Tuomainen, J., Wolfsthurn, S. V. G. Z., Von Grebmer Zu Wolfsthurn, S., Kohut, Z., Kulakova, E., Meziere, D., Politzer-Ahles, S., Rousselet, G., Rueschemeyer, S.-A. A., Segaert, K., Tuomainen, J. & Wolfsthurn, S. V. G. Z. (2019). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *bioRxiv*, *28*(10), 1–29

Nieuwland, M. S. & Martin, A. E. (2017). Neural Oscillations and a Nascent Corticohippocampal Theory of Reference. *Journal of Cognitive Neuroscience*, *29*(5), 896–910

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., Donaldson, D. I., Kohút, Z., Rueschemeyer, S. A. & Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, 1–24

Nieuwland, M. S. & Van Berkum, J. J. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, *24*(3), 691–701

Nieuwland, M. S. & Van Berkum, J. J. a. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of cognitive neuroscience*, *18*(7), 1098–1111

Norris, D., McQueen, J. M. & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4–18

Oaksford, M. & Chater, N. (1994). A Rational Analysis ofthe Selection Task as Optimal Data Selection. *Psychological Review*, *101*(4), 608–631.

Oaksford, M. & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review*, pp. 381–391.

Oaksford, M. & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in cognitive sciences*, *5*(8), 349–357

Oaksford, M., Chater, N. & M., O. (2009). Precis of of Bayesian Rationality: The probabilistic approach to human reasoning. *The Behavioral and brain sciences*, *32*(1), 69–120

Oliva, A. (2005). Gist of the scene. *Neurobiology of Attention*, pp. 251–256.

Oliva, A. & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision2*, *42*(3), 145–175

Osterhout, L. & Holcomb, P. J. (1992). Event-related potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806.

Palmer, S. E. (1999). *Vision Science*. Cambridge, MA.: MIT Press.

Parikh, P. (2001). *The use of language*. Stanford, CA.: CSLI Publications: Stanford University.

Partee, B. H. (1984). Compositionality. In F. Veltman (ed.), *Varieties of Formal Semantics* pp. 281–312. Dordrecht: Foris.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, *95*(1), e41–e74

Payne, B. & Federmeier, K. D. (2019). Individual Differences in Reading Speed are Linked to Variability in the Processing of Lexical and Contextual Information: Evidence from Single-trial Event-related Brain Potentials. *Word*, *65*(4), 252–272

Payne, B. R. & Federmeier, K. D. (2018). Contextual constraints on lexico-semantic processing in aging: Evidence from single-word event-related brain potentials. *Brain Research*, *1687*, 117–128

Payne, B. R., Lee, C.-l. & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology*, *52*(11), 1456–1469

Pelletier, F. J. (1994). The Principle of Semantic Compositionality. *Topoi*, *13*, 11–24.

Pelletier, F. J. (2012). Holism and compositionality. *The Oxford handbook of compositionality*, pp. 149–175.

Pennington, J., Socher, R. & Manning, C. D. (2014). GloVe : Global Vectors for Word Representation.

Petitot, J. (1988). Morphodynamics and the categorical perception of phonological units. *Theoretical Linguistics*, *15*(1-2), 61–80

Phillips, C. (2003). Linear Order and Constituency. *Linguistic Inquiry*, *34*(1), 37–90

Phillips, S. (2018). Going beyond the data as the patching (sheaving) of local knowledge. *Frontiers in Psychology*, *9*(OCT)

Phillips, S. (2020). Sheaving-a universal construction for semantic compositionality. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *375*(1791), 20190303

Piantadosi, S. T., Tily, H. & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291

Pickering, M. J. & Garrod, S. (2013a). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(04), 329–347

Pickering, M. J. & Garrod, S. (2013b). How tightly are production and comprehension interwoven? *Frontiers in Psychology*, *4*(April), 1–2

Plate, T. (2006). Distributed Representations. In *Encyclopedia of Cognitive Science* pp. 1–15. Chichester: John Wiley & Sons, Ltd

Poeppel, D. (2012). The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology*, *29*(1-2), 34–55

Potts, C. (2004). *The Logic of Conventional Implicatures*. Oxford University Press

Prashant Parikh (2000). Communication, Meaning, and Interpretation. *Linguistics and Philosophy*, *23*, 185–212.

Purver, M. (2015). From Distributional Semantics to Distributional Pragmatics ? In *Proceedings of the IWCS 2015 Workshop on Interactive Meaning Construction*, Lisbon, Portugal.

Purver, M. & Kempson, R. (2004). Incremental parsing, or incremental grammar? In *Proceedings of the Workshop on Incremental Parsing Bringing Engineering and Cognition Together - IncrementParsing '04*, pp. 74–81., Morristown, NJ, USA. Association for Computational Linguistics

Quante, L., Bölte, J. & Zwitserlood, P. (2018). Dissociating predictability, plausibility and possibility of sentence continuations in reading: evidence from late-positivity ERPs. *PeerJ*, *6*, e5717

Rabovsky, M. (2019). Implicit semantic prediction error can account for N400 effects on articles that do not differ in meaning: A neural network model. *bioRxiv*

Rabovsky, M. (2020). Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model. *Neuropsychologia*, *143*, 107466

Rabovsky, M., Hansen, S. S. & Mcclelland, J. L. (2016). N400 amplitudes reflect change in a probabilistic representation of meaning: Evidence from a connectionist model. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, pp. 2045–2050.

Rabovsky, M., Hansen, S. S. & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705

Rabovsky, M. & McClelland, J. L. (2020). Quasi-compositional mapping from form to meaning: a neural network-based approach to capturing neural responses during human language comprehension. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *375*(1791), 20190313

Rabovsky, M. & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition*, *132*(1), 68–89

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J. & Hochreiter, S. (2020). Hopfield networks is all you need. *arXiv*.

Rauss, K. & Pourtois, G. (2013). What is bottom-up and what is top-down in predictive coding. *Frontiers in Psychology*, *4*(MAY), 1–8

Reboul, A. (2011). A relevance theoretic account of the evolution of implicit communication. *Studies in Pragmatics*, *13*, 1–21.

Recanati, F. (2002). Does Linguistic Communication Rest on Inference? *Mind and Language*, *17*(1&2), 105–126.

Recanati, F. (2003). What is said and the semantics/pragmatics distinction. In C. Bianchi (ed.), *The Semantics/Pragmatics Distinction*. CSLI Publications: Stanford University

Recanati, F. (2005). Pragmatics and Semantics. In *The Handbook of Pragmatics* pp. 864. Wiley-Blackwell.

Recanati, F. (2017). Local pragmatics: reply to Mandy Simons. *Inquiry (United Kingdom)*, *60*(5), 493–508

Reddy, M. J. (1979). The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony (ed.), *Metaphor and Thought* pp. 284–310. Cambridge, MA.: Cambridge University Press.

Riest, C., Jorschick, A. B. & de Ruiter, J. P. (2015). Anticipation in turn-taking: Mechanisms and information sources. *Frontiers in Psychology*, *6*(FEB), 1–14

Rohde, H. & Horton, W. S. (2014). Anticipatory looks reveal expectations about discourse relations. *Cognition*, *133*(3), 667–691

Rohde, H. & Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, *29*(8), 912–927

Rohde, H., Levy, R. & Kehler, A. (2011). Anticipating explanations in relative clause processing. *Cognition*, *118*(3), 339–358

Roland, D., Yun, H., Koenig, J. P. & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition*, *122*(3), 267–279

Rommers, J. & Federmeier, K. D. (2018). Predictability's aftermath: Downstream consequences of word predictability as revealed by repetition effects. *Cortex*, *101*, 16–30

Rosa, E. C. & Arnold, J. E. (2017). Predictability affects production: Thematic roles can affect reference form selection. *Journal of Memory and Language*, *94*, 43–60

Rosch, E. & Lloyd, B. L. (1978). *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum.

Rosen, K. H. (2012). *Discrete Mathematics and Its Applications*. McGraw-Hill Science/Engineering/Math.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408

Rotaru, A. S., Vigliocco, G. & Frank, S. L. (2018). Modeling the Structure and Dynamics of Semantic Processing. *Cognitive Science*, *42*(8), 2890–2917

Roth, E. M. & Shoben, E. J. (1983). The effect of context on the structure of categories. *Cognitive Psychology*, *15*(3), 346–378

Russell, B. (1903). *The principles of mathematics*. London: Norton.

Russell, B. (1910). On the Nature of Truth and Falsehood. In B. Russell (ed.), *Philosophical Essayes*. Longmans, Green.

Russell, B. (2012). *Probabilistic reasoning and the computation of scalar implicatures*. Brown University, Providence, Rhode Island.

Ryskin, R., Levy, R. P. & Fedorenko, E. (2020). Do domain-general executive resources play a role in linguistic prediction? Re-evaluation of the evidence and a path forward. *Neuropsychologia*, *136*(November 2019), 107258

Sadrzadeh, M. (2016). Quantifier Scope in Categorical Compositional Distributional Semantics. *Electronic Proceedings in Theoretical Computer Science*, *221*, 49–57

Sadrzadeh, M., Clark, S. & Coecke, B. (2013). The Frobenius anatomy of word meanings I: Subject and object relative pronouns. *Journal of Logic and Computation*, *23*(6), 1293–1317

Sadrzadeh, M., Kartsaklis, D., Balkir, E., Kartsaklis, D. & Sadrzadeh, M. (2018). Sentence Entailment in Compositional Distributional Semantics. *Annals of Mathematics and Artificial Intelligence*

Sassenhagen, J. & Fiebach, C. J. (2019). Finding the P3 in the P600: Decoding shared neural mechanisms of responses to syntactic violations and oddball targets. *NeuroImage*, *200*, 425–436

Saul, J. M. (2002). What is said and psychological reality; Grice's project and relevance theorists' criticisms. *Linguistics and Philosophy*, *25*, 347–372.

Scholman, M. C., Rohde, H. & Demberg, V. (2017). "On the one hand" as a cue to anticipate upcoming discourse structure. *Journal of Memory and Language*, *97*, 47–60

Schuster, S., Chen, Y. & Degen, J. (2020). Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5387–5403., Stroudsburg, PA, USA. Association for Computational Linguistics

Scontras, G., Tessler, M. H. & Franke, M. (2018). Probabilistic language understanding: An introduction to the Rational Speech Act framework. Retrieved on 2021-01-12 from `https://www.problang.org`

Sedivy, J. C., K. Tanenhaus, M., Chambers, C. G. & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147

Sennet, A. (2016). Ambiguity. In E. N. Zalta (ed.), *The {Stanford} Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, (spring 201 éd.).

Shetreet, E., Alexander, E. J., Romoli, J., Chierchia, G. & Kuperberg, G. (2019). What we know about knowing: Presuppositions generated by factive verbs influence downstream neural processing. *Cognition*, *184*(December 2017), 96–106

Skyrms, B. (2010). *Signals*. Oxford University Press

Smith, J. M. & Price, G. R. (1973). The Logic of Animal Conflict. *Nature*, *246*(5427), 15–18

Smith, N. A. (2019). Contextual Word Representations: A Contextual Introduction. *arXiv*, pp. 1–15.

Smith, N. J. & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Conference*, pp. 1637–1642.

Smith, N. J. & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319

Socher, R., Bauer, J., Manning, C. D. & Ng, A. Y. (2013). Parsing with compositional vector grammars. *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, *1*, 455–465.

Socher, R., Huang, E. & Pennington, J. (2011). Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. *Advances in Neural Information Processing Systems*, pp. 801–809.

Sperber, D., Clément, F., Heintz, C., Mercier, H., Origgi, G., Wilson, D., Mascaro, O., Mercier, H., Origgi, G. & Wilson, D. (2010). Epistemic Vigilance. *Mind and Language*, *25*(4), 359–393

Sperber, D. & Wilson, D. (1995). *Relevance: Communication and Cognition*, volume 10. Oxford, UK: Blackwell Publishers.

Sperber, D. & Wilson, D. (1996). Fodor's Frame Problem and Relevance Theory. *Behavioral and Brain Sciences*, *19*(3), 530–532.

Sperber, D. & Wilson, D. (2002). Pragmatics, Modularity and Mind-reading. *Mind and Language*, *17*(April), 3–23

Sperber, D. & Wilson, D. (2012). A deflationary account of metaphors. In R. W. J. Gibbs (ed.), *The Cambridge Handbook of Metaphor and Thought* pp. 84–106. Cambridge: Cambridge University Press

St. John, M. F. & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*(1-2), 217–257

Stalnaker, R. (1976). Propositions. In A. MacKay & D. D. Merrill (eds.), *Issues in the Philosophy of Language* pp. 79–91. New Haven: Yale University Press.

Stalnaker, R. (2006). Saying and Meaning, Cheap Talk and Credibility. In A. Benz, G. Jäger, & R. van Rooij (eds.), *Game Theory and Pragmatics* pp. 83–100. London: Palgrave Macmillan UK

Stanojević, M. & Steedman, M. (2019). CCG Parsing Algorithm with Incremental Tree Rotation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 228–239.

Staub, A. (2015). The Effect of Lexical Predictability on Eye Movements in Reading: Critical Review and Theoretical Interpretation. *Language and Linguistics Compass*, *9*(8), 311–327

Staub, A., Grant, M., Astheimer, L. & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1–17

Steedman, M. (1996). A very short introduction to CCG. *Unpublished paper. http://www. coqsci. ed. ac. uk/steedman/paper. html*, pp. 1–8.

Steedman, M. (1999). Alternating Quantifer Scope in CCG. In *Conference: 27th Annual Meeting of the Association for Computational Linguistics*, pp. 301–308.

Steedman, M. (2014). Categorial Grammar. In A. Carnie, Y. Sato, & D. Siddiqui (eds.), *The Routledge Handbook of Syntax* pp. 670–701. Routledge

Steedman, M. (2019). Combinatory Categorial Grammar. In A. Kertész, E. Moravcsik, & C. Rákosi (eds.), *Current Approaches to Syntax - A Comparative Handbook*. de Gruyter Mouton.

Steedman, M. & Baldridge, J. (2011). Combinatory Categorial Grammar. In R. D. Borsley & K. Börjars (eds.), *Non-Transformational Syntax* pp. 181–224. Oxford, UK: Wiley-Blackwell

Steyvers, M. & Griffiths, T. (2010). Probalistic Topic Models. *Latent Semantic Analysis: A Road To Meaning*, *3*(3), 993–1022

Strasser, C. & Antonelli, G. A. (2019). Non-monotonic Logic. Retrieved on 2021-01-14 from `https://plato.stanford.edu/archives/sum2019/entries/logic-nonmonotonic/`

Szewczyk, J. M. & Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language*, *68*(4), 297–314

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, *30*(4), 415–433

Tenenbaum, J. B., Griffiths, T. L. & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318

Tessler, M. & Goodman, N. D. (2014). Some arguments are probably valid: Syllogistic reasoning as communication. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*, 1574–1579

Torralba, A. & Sinha, P. (2001). Statistical context priming for object detection. In *Proceedings of the International Conference on Computer Vision*, pp. 763–770., Vancouver, Canada

Trapp, S., Lepsien, J., Kotz, S. A. & Bar, M. (2016). Prior probability modulates anticipatory activity in category-specific areas. *Cognitive, Affective and Behavioral Neuroscience*, *16*(1), 135–144

Turney, P. D. & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188

Van Berkum, J. J. & Nieuwland, M. S. (2019). A Cognitive Neuroscience Perspective on Language Comprehension in Context. In *Human language: From genes and brains to behavior* pp. 429–442. Cambridge, MA: MIT Press.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V. & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of experimental psychology. Learning, memory, and cognition*, *31*(3), 443–467

van Dijk, T. A. & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

van Eijck, J. (2001). Incremental dynamics. *Journal of Logic, Language and Information*, *10*(3), 319–351

van Eijck, J. & Kamp, H. (1997). Representing Discourse in Context*. In *Handbook of Logic and Language* pp. 179–237. Elsevier

van Eijck, J. & Kamp, H. (2011). Discourse Representation in Context. *Handbook of Logic and Language*, pp. 181–252.

Van Petten, C. & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190

van Rooij, R. (2004). Signaling Games Select Horn Strategies. *Linguistics and Philosophy*, *27*(4), 493–527.

van Rooy, R. (2004). Signalling Games Select Horn Strategies. *Linguistics and Philosophy*, *27*(4), 493–527

Vandepitte, S. (2001). Anticipation in Conference Interpreting : A Cognitive Process. *Revista Alicante de Estudios Ingleses*, *14*, 323–335.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009.

Veldre, A. & Andrews, S. (2018). Beyond cloze probability: Parafoveal processing of semantic and syntactic information during reading. *Journal of Memory and Language*, *100*, 1–17

Venhuizen, N. J., Bos, J., Hendriks, P. & Brouwer, H. (2018). Discourse semantics with information structure. *Journal of Semantics*, *35*(1), 127–169

Venhuizen, N. J., Crocker, M. W. & Brouwer, H. (2019). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, *56*(3), 229–255

Wang, C., Paisley, J. & Blei, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. *Journal of Machine Lea*, *15*, 752–760

Warren, T. & McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin & Review*, *14*(4), 770–775

Wilson, D. (2016). Relevance Theory. In Y. Huang (ed.), *The Oxford Handbook of Pragmatics* pp. 57–98. Oxford University Press

Wilson, D. & Carston, R. (2007). A Unitary Approach to Lexical Pragmatics : Relevance , Inference and Ad Hoc Concepts. In N. Burton-Roberts (ed.), *Pragmatics* pp. 230–259. London: Palgrave.

Wilson, D. & Sperber, D. (2012). Truthfulness and Relevance. In *Meaning and Relevance* pp. 47–83. Cambridge University Press.

Wilson, D. & Sperber, D. A. N. (2004). Relevance Theory. In L. Horn & G. Ward (eds.), *Handbook of Pragmatics* pp. 607–632. Blackwell Publishers.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell.

Xu, F. & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272

Yun, H., Mauner, G., Roland, D. & Koenig, J.-P. (2012). The Effect of Semantic Similarity i s a Function of Contextual Constraint Experiment to Generate Reading Times. In *Proceedings of the 34th Conference of the Cognitive Science Society (CogSci 2012)*, pp. 1191–1196.

Zeevat, H. (2015a). Perspective on Bayesian Natural Language Semantics and Pragmatics. In H. Zeevat & H.-C. Schmitz (eds.), *Bayesian Natural Language Semantics and Pragmatics*, volume 2 of *Language, Cognition, and Mind* pp. 1–24. Cham: Springer International Publishing

Zeevat, H. (2015b). Perspectives on Bayesian Natural Language Semantics and Pragmatics. In H. Zeevat & H.-C. Schmitz (eds.), *Bayesian Natural Language Semantics and Pragmatics* pp. 1–24. Springer International Publishing

Zemel, R., Dayan, P. & Pouget, A. (2008). Probabilistic Interpretation of Population Codes Communicated by Terrence Sanger. *Neural Computation*, *10*(2), 403–430

Zufferey, S. (2016). Pragmatic acquisition. In *Handbook of Pragmatics*. Amsterdam: John Benjamins Publishing Company

Zwaan, R. A. (2008). Time in Language, Situation Models, and Mental Simulations. In P. Indefrey & M. Gullberg (eds.), *Time to Speak: Cognitive and Neural Prerequisites for Time in Language*. Language Learning Research Club.

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin and Review*, *23*(4), 1028–1034

Zwaan, R. A. & Radvansky, G. A. (1998). Situation Models in Language Comprehension and Memory. *Psychological Bulletin*, *123*(2), 162–185