Focus in Communication: An Information-Theoretic Approach

Author:
Philip Schulz
Student ID: 10408835

1st Supervisor:
Dr. Michael Ramscar

2nd Supervisor:
Dr. Christian Ebert

Date of Submission:
October 24, 2013

Philip Schulz
Conventions

In order to make this thesis more legible I tried to follow some writing conventions throughout. If words are used as objects of presentation they are italicized. Technical terms are marked by sans-serif font on the first mention. Finally, since emphasizing in \LaTeX is very akin to italics, I decided to use underlining instead.

The thesis is structured in chapters. The numbering of examples and footnotes begins anew with each chapter. Definitions are numbered in ascending fashion throughout the thesis. Figures are labeled with the number of the chapter they occur in followed by their position relative to other figures within that chapter. The same holds for equations.

The bibliography at the end of this thesis is structured alphabetically, with the year of publication occurring as the last item within a reference.
Acknowledgements

Finally the time has come to write up the acknowledgements for this piece of work. This for sure is one of the parts I have been looking forward to most. Let me start out by thanking everyone at the Seminar für Sprachwissenschaft at the University of Tübingen for the excellent education they provided me with. After having studied at two other institutions that are widely regarded as top schools in Europe and the world, I can safely say that I benefited most from the teaching at the SFS. Thank you guys for that!

I would also like to thank a couple of people in particular. First of all there are my three linguistic mentors, which I will mention in the order in which I got to know them. Wolfgang Sternefeld instilled a critical attitude in me that makes me question every theory I come across nowadays. Being able to critically assess other people’s work instead of blindly following well-presented ideas has proven to be a cornerstone in my life as a linguist and researcher. Wolfgang served me as role model in this respect and others. Yi Xu from the University College London taught me virtually everything I know about phonetics. If I can make any sense of what focus does in speech, it’s because of him. Also, he taught me that language, and speech in particular, is incredibly rich and should not be constrained to some formal system just to make it look elegant. Yi also let me use the phonetic labs at UCL to conduct my own research and provided my with valuable advise on how to design and run experiments. Finally, there is Michael Ramscar. Michael gave a talk at the SFS about a year and a half ago in which he pinned down a lot of things that had been rumoring in me for quite a while. Immediately afterwards I asked him whether he would supervise this thesis (luckily he agreed). Michael helped me expand my knowledge about language and also opened my eyes for a completely new and compellingly simple view on language. Not only did this help me to overcome my frustration with stipulative nature of traditional linguistics, it also made me understand and appreciate precisely formulated computational models which should constitute the basis of all linguistic theories. The fact that I am now more and more moving into statistically driven fields of linguistics is largely his doing. Besides being great mentors, all three of them are great personalities with very special and unique characters, who I had a lot of fun discussing with.

Although they make you testify that your Master’s thesis is exclusively your own work, it never really is (how could it be if it is meant to comprise 5 years of education?). At this point I would like to thank the people how supported me practically and morally while I was writing my thesis and in the time leading up to that. Christian Ebert has volunteered to read and grade my thesis for the second time, after he already supervised my Bachelor’s thesis. I can’t thank him enough for that since without an examiner there would be no degree (sorry that this thesis ended up being so long). Christian also was approachable whenever needed and helped me circumvent the cliffs of academic bureaucracy. Fabian Tomaschek also helped me out for the second time. After giving me an Anova crash course for my Bachelor’s thesis he took recordings for my experiment this time around. I can’t thank him enough for that as otherwise this thesis would
likely not report an experiment. The last one on the practical side is Evante Garza-Licunde. Evante wrote me a javascript that finally got said experiment running.

Detmar Meurers and his team offered me a place in which to work and learn for more than two years. Detmar’s enthusiasm and the cosy atmosphere he managed to create in his team made my time in Tübingen definitely more pleasant. And, yes, the pay obviously helped a lot as well. When it came to clearing my head, exercising a bit and occasionally being punched in the face, Stephan Ripple and his Ving Tsun group in Pfullingen were indispensable. I had a great time with you guys and miss training with you a lot. The working routine in Tübingen was otherwise interrupted by 3 to 4 daily coffee breaks, most of which I enjoyed in the company of Marisa, Niko and Sarah. Somewhat surprisingly, we all finished in time despite constantly frequenting the bakery around the corner. Finally, I want to thank my amazing friends in London and Berlin who were always there for me when I needed them. They also got my head out of the academic clouds on occasion and put my feet back on the ground. At uni I learned a lot of stuff, but with them I learned about life.

Natürlich möchte ich mich an dieser Stelle auch herzlich bei meinen Eltern bedanken, die mich immer unterstützt und mich dazu angehalten haben, meine Pläne zu verwirklichen. Dies tun sie auch heute noch, obwohl sie wahrscheinlich immer noch nicht ganz verstehen, was ich da eigentlich so die ganze Zeit mache. Vielen Dank dafür!
Abstract

This thesis presents a new approach to focus, and information structure more generally, based on information theory. Information theory provides us with a notion of information that is purely quantitative and therefore relatively easy to measure empirically. I show that a prominent feature of focus marking, namely that it leads to words being pronounced longer, can be directly tied to insights about language processing that are based on information theory and the notion of predictability that it employs. The theory advocated here is that focus falls on relatively unpredictable words so as to allow the speaker more time for lexical access but also to keep the rate of information transmitted constant. This is in line with the two guiding principles of the research reported here. The first one is that one should only assume what one already knows from prior research. This clearly restricts the room for speculation. The second principle is to try to explain empirical observations with simple and domain-general mechanisms whenever possible. In the present work this is achieved through connecting my proposal to a general theory of behavioural learning. Specific notions of grammar and language-specific processing become therefore obsolete for our purposes.

A second contribution of this thesis is that it explicitly presents language as a communicative and hence social phenomenon. Applied to focus this means that it not only serves as a linguistic device that helps to keep the flow of information constant but also a social device. I report an experiment in which this hypothesis was tested. We employed crowdsourcing on a large scale in order to obtain friendliness and other social judgements for question-answer dialogues in which focus placement had been systematically manipulated. I show that focus has a strong effect on these judgements. Subsequently, I make the argument that even this social impact of focus can ultimately be tied to the information-theoretic notion of uncertainty reduction.

In sum this thesis offers a new view on focus according to which focus serves to keep the stream of information stable and to reduce uncertainty about the speaker’s intentions. This in turn makes focus an important linguistic and social device.
Zusammenfassung


Zusammenfassend eröffnet diese Arbeit einen neuen Blick auf Fokus, folgern dem Fokus dazu dient, den Informationsfluss stabil zu halten und die Ungewissheit über die Intentionen des Sprechers zu reduzieren. Dadurch wird Fokus zu einem wichtigen linguistischen und sozialen Medium.
# Contents

1 Introduction 1

2 Focus: Form and Function 7
   2.1 Form ........................................ 7
   2.1.1 Phonetics ................................. 7
   2.1.2 Acquisition ............................... 12
   2.1.3 Syntax .................................... 18
   2.2 Function ................................... 20
   2.2.1 Views from Semantics ................... 20
   2.2.2 Views from Psycholinguistics .......... 22

3 Language and Information 31
   3.1 Basic Notions of Information Theory ........ 34
   3.1.1 Estimation ................................. 41
   3.2 Is This Even Real? ........................... 44
   3.3 Learning ..................................... 53
   3.4 Additional remarks .......................... 60
       3.4.1 A Note on Grammar ..................... 60
       3.4.2 A Note on Semantics ................... 61

4 Focus and Information 63
   4.1 Focus Duration and Information ............ 63
   4.2 Focus as Social Behaviour .................. 69
       4.2.1 Method .................................. 70
       4.2.2 Analysis and Results ................... 72

5 Conclusion and Outlook 81

Bibliography 85
Chapter 1

Introduction

Focus has been one of the central notions in linguistics for decades, starting at least from Halliday (1967). The reason is that focus is to a large extent encoded by phonetic means in most European languages and hence also in English. Since English has been the primary object of study for most of the research done in linguistics, focus has been seen as the missing link between word or sentential meaning and the meaning conveyed by other components of speech, such as intonation, syllable duration and the like. Focus incorporates precisely these aspects of speech as it affects a word’s duration, pitch range and contour and, in some cases, intensity (i.e. loudness as measured through changes in air pressure). The interest in focus was further enhanced with the advent of information structure research (Chafe, 1976). The idea put forward by scholars interested in information structure is that words and their associated concepts may receive different levels of activation depending on how they are uttered and which position in an utterance they occupy. As a consequence of that activation, words may have different impact on a conversation depending on their information status. Naturally, information structure research has adopted focus as one of its main objects of investigation since focus is clearly detectable due to its phonetic distinctiveness. Moreover, focus has been associated with many non-lexical meaning components and hence has been seen as a prime example manifesting the claims put forward by information structure research. For the same reason it has also evolved into a major topic in semantics, starting from Jackendoff (1972) and culminating in a now standard theory presented in Rooth (1985, 1992). However, due to the ever-changing nature of approaches to information structure and grammar, which many information structural theories consider themselves to be a part of (e.g. Rochemont, 1986; Selkirk, 1995; Hartmann and Zimmermann, 2009; Winkler, in press), as of yet there is no consensus about which phenomena should be classified as focus. Moreover, the notion of meaning that the afore-mentioned accounts of focus employ is arguably not very well suited to model the communicative impact of focus. Following the tradition of Frege (1892), meaning is seen as a logical notion under which sentences can be either true or false. Subsentential parts of utterances, which focus is an exam-
ple of, only contribute to the truth or falsity of a sentence. Social aspects of communication, such as politeness, flirtation, negotiation of one’s social rank and many more cannot be captured under this conception of meaning.

In parallel, focus has attracted some attention in the field of phonetics. It turned out to be fairly easy to consistently elicit focus in question-answer contexts. This observation led to a series of experiments in the mid-80’s (Cooper et al., 1985; Eady and Cooper, 1986). Building upon some of the communicative effects of focus, they provided very detailed insights into how focus is encoded in speech. Moreover, since the experiments were carefully designed to (i) consistently elicit focus and (ii) exclude other intervening effects on the speech produced by the speaker as far as possible, the authors had to define focus very precisely. The focused word in these experiments is always contained in an answer. More precisely it is the one that addresses the wh-element in the question. This rather narrow definition of focus has many advantages. As already pointed out, it allows for simply designed yet efficient experiments. Moreover it removes some of the explanatory burden that any theory faces which views focus in much broader terms. Such a theory would likely have to deal with much more variant data. Most importantly, however, this definition of focus is extensible in the sense that any theory of focus will have to include question-answer pairs and therefore can be built on top of this definition. This is why, in spite of its imperfection, I will adopt it as a working definition throughout this thesis.

**Definition 1 (Focus).** Focus is the part of the answer that addresses the an explicit question.

In the case of wh-questions this means that the focused word in the answer will most likely correspond to the wh-element in the question. For yes/no-questions it means that the focus will most likely be either *yes* or *no*. An example for wh-questions is given in (1), where *carrots* is most likely focused in the answer.

(1) A: What did you have for lunch yesterday?  
B: I had carrots for lunch yesterday.

Notice that the definition only talks about explicitly asked question and therefore excludes any sort of implicit questions such as Questions Under Discussion (Roberts, 2012). The reason simply is that implicit questions are hard to detect (no algorithm for doing so has yet been given) and hence no empirically justifiable theory can be build on them.

Moreover, this definition is to be set apart from others based on either alternative semantics or purely phonetic features of focus. The de facto standard

---

1Whenever an explicit question is given in my examples I will refrain from the standard practice of putting the focus in uppercase. The reason is that this notation is highly suggestive and may lead to wrong intuitions about focus. It often happens that authors uppercase alleged foci in text but completely neglect its phonetic properties in speech. Moreover, I think that any phenomenon should always be readily identifiable from its definition without any need for further marking.
definition of focus\textsuperscript{2} in semantics is spelled out in Krifka (2008). He states that “Focus indicates the presence of alternatives that are relevant for the interpretation of linguistic expressions” (Krifka, 2008, Section 3.1). The problem with this definition is that it is not general enough but solely based on a particular semantic theory of focus. Worse even, it does not even define what “alternatives” are supposed to be. As it turns out, alternatives are being looked for all the time during language comprehension (see e.g. Tanenhaus et al., 1995; Dahan et al., 2002). However, there seem to be no alternatives that are in any way particular to focus (Gernsbacher and Jescheniak, 1995). Finally, Krifka’s conception of focus is not even a definition but merely a description of (alleged) effects. I therefore conclude this conception cannot be adequate.

A second line of focus definitions comes from phonology. There, focus is often conceived of as some form of intonational prominence or highlighting (see e.g. Rooth, 1985, p.1). The problem with this kind of definition is that, again, there is consensus as to how to define prominence. This is in fact a long-standing issue in prominence research (Arnold et al., 2012). Moreover, there are many other functions of speech that are not related to focus but also involve prominence, such as emotion (Chuenwattanapranithi et al., 2008) or marking of importance (Watson et al., 2008). Hence, focus cannot be delineated from other phenomena just on the basis of its phonetics. Therefore, this line of definitions has to be dismissed, as well.

Finally, there are pragmatic definitions that equate focus with new information (Ward and Birner, 2004, p.157). These can however be refuted easily, as pointed out by Schwarzschild (1999). Consider example (2).

\begin{example}
A: Would you like cheese or cake? B: I’d like cheese, please.
\end{example}

Here, \textit{cheese} will be the focus in B’s answer. However, it is by no means new. Hence, we also have to close off this attempt of a focus definition. In sum, it seems that def. 1, though not perfect, is the most adequate definition of focus.

\section*{Thesis Goal and Structure}

The goal of this thesis is to unite ideas about the communicative function of focus with findings about its form (i.e. its phonetic realization). In doing so, I will emphasize the communicative significance of focus from a broader communication-centered perspective. While certain ideas from formal linguistics will be adopted in the process, I will part with traditional approaches to focus in several important aspects. First of all, I will look at communication as a form of human behaviour. Language in turn will be viewed as a specific form of communication. This has some direct implications for how I will approach language from a theoretical perspective. I will treat language on par with other forms of human behaviour and with other forms of communication.

\textsuperscript{2}As of October 1st 2013, this paper has received 314 cites according to google scholar. In the field of linguistics, and especially in semantics, this can be considered very high impact, indicating that the definitions from this paper are widely accepted.
In addition, I will not assume any language production or processing faculties or mechanisms other than those that are obvious (e.g. visible articulators such lips, tongue and teeth) or those that have been experimentally observed. Moreover, no claim will be made that any of the faculties involved in language production and comprehension be specific to language. This is because to the best of my knowledge no convincing evidence has yet been provided that such a faculty exists. On the contrary, it seems that most of the faculties that are crucial for language are also operational in other domains. For example, many articulators are indispensable for eating and breathing. Our ears gather valuable information about our environment from signals that are non-linguistic in nature. An example is the noise of an approaching car, the perception of which may save us from being run over. On a neural level it has been shown by Hertrich et al. (2009) that people who are born blind can allocate a great portion of their visual cortex to language processing. During speech comprehension they show greater neural activity in the visual cortex than sighted people. This in turn allows them to process and understand speech that has been accelerated to up to 32 times the normal speaking rate. Sighted people by and large lack this enhanced processing ability, indicating that blind people do indeed have a larger processing ability for speech. Furthermore it has been shown that if the parts of the brain that are usually associated with language are lesioned at an early age, other areas of the brain respond more strongly to language than they would in healthy subjects. Hence there are arguments for (i) a possible expansion of language-related areas in the brain if resources are available and (ii) the re-allocation of neural resources to language processing and production in case certain brain areas are damaged. This makes it indeed hard to believe that any part of the brain be uniquely pre-destined to host faculties that are necessary for the use of language.

The second way in which I will depart from conceived wisdom in formal linguistics is the treatment of meaning. As shortly pointed out above, the traditional logic-centered view of linguistic meaning restricts the possible coverage of any theory that builds on that view. Consequently, this notion of meaning does not fit well with the goal of examining focus from a communicative-functional perspective. In fact, any fixed notion of meaning would most likely unnecessarily confine the space of communicative aspects of language that one can describe. This may in fact be a benefit for practical applications, where it is often helpful to confine the search space of possible meanings. An example at hand is document retrieval and classification. Here, vector space models of word meaning have yielded great improvements (see e.g. Widdows, 2004). Nevertheless, in accordance with the approach to language as a form of human behaviour, no such concrete notion of meaning will be employed here since we simply do not know which meaning components are crucial for human understanding of others’ behaviour. Rather, I will assume that linguistic behaviour can be observed just like other sensorily detectable events and consequently reduce an observer’s uncertainty about his surroundings. This initial description of the present approach to language is held deliberately vague, so as to emphasize the parallels between language and other observable events such as smells and move-
ments. The main difference between these and language is that language is for the most part used consciously whereas we hardly have conscious control over the body smells we emit and only limited conscious control over our movements. I will return to a more precise description of language as a human behaviour and how it can be understood using concepts from information theory in Chapter 3. Also, I will connect these insights to a theory of learning in Section 3.3.

In line with the view of language as a communicative phenomenon, the research reported here concentrates on two things. First, I will show that focus can be integrated with what we know about general language processing. In particular, if language is viewed from an information-theoretic angle, some of the distributional and functional properties of focus can be explained straightforwardly. I elaborate on this proposal in Section 4.1. After that I will turn to the social functions that focus may have and report the results of a large-scale online study investigating the effects of focus on discourse coordination and how confidence, sincerity and friendliness ascriptions vary as a function of focus matching in question-answer contexts. These results are reported in Section 4.2.

Before I turn to the information-theoretically motivated view on focus I will first review results concerning the phonetic realization of focus in languages such as English in Section 2.1. These results are backed up by a discussion about the learnability of focus. I then go on to discuss research about the function of focus from a partly semantic but most psycholinguistic perspective in Section 2.2.
Chapter 2

Focus: Form and Function

This chapter is dedicated to the introduction of a solid background knowledge that will be needed to understand the experiments in Section 4.2. Moreover, it also serves as a literature review, that summarizes results and ideas about focus from phonetics, formal linguistics and psycholinguistics and psychology. At the same time, the state of the art in focus research will be assessed critically.

The title of this chapter already hints at the diverse nature of the phenomenon of focus. Focus marking is very rich, including components from both phonetics and morphosyntax. Hence, it is interesting in its own right to look at focus marking devices and figure out how these work. However, one should not neglect that focus has very important communicative functions. An ideal theory of focus would of course attempt to tie the formal and functional aspects of focus together and explain establish a correspondence between form and function. The present chapter attempts to make the link between form and function whenever possible and also introduces new functional aspects of focus that are neglected often neglected in linguistic research.

2.1 Form

2.1.1 Phonetics

I will now turn to the phonetic characteristics of focus. Since English is the language we investigated in our experiments, it is necessary to explain the phonetic encoding mechanisms that are used by speakers of that language to signal focus. Most prominently, focus is associated with a rise in pitch, i.e. height of voice. In particular, focus is often equated with “pitch accent” (Selkirk, 1995). While it is certainly true that pitch is an important part of focus coding, it is not so much the absolute height of the pitch that signals focus, but rather the slope of the pitch contour. It is well known that pitch varies between speakers, in particular if they are of different gender (see Xu et al., 2013, and references therein). Furthermore, pitch height and contour also signals body size and
Figure 2.1: Focus on three different locations in a Mandarin statement. The curves are averaged over 4 speakers each of which produced 24 repetitions; Hz values are shown on the y-axis. Graphic taken from Xu (2005).
encodes emotions like happiness and anger (Chuenwattanapranithi et al., 2008). Therefore, pitch height (or “accent”) cannot be enough for identifying a focus.

**Statements**

In order to understand how the pitch contour of statements is affected by focus, let us take a look at Figure 2.1. Although the Figure shows contours of Mandarin speakers, the observations we make also apply to English as these two languages code focus in similar ways.

The most obvious impact of focus on pitch is that the pitch range gets expanded. Note that focus only amplifies the pitch contour and does not change its directionality. Hence, it is more appropriate to call this a pitch range expansion than accenting. It also appears that this pitch range expansion is stronger in sentence-initial and sentence-medial position than in sentence-final position. In fact, this observation is consistent with empirical findings. The pitch range expansion usually is at most half as wide for a final focus than for foci in other positions (Cooper et al., 1985; Eady and Cooper, 1986; Xu, 2005). Although the pitch height of a final focus does differ significantly from the pitch in the same position of a sentence that does not contain any focus (the *None* contour in Figure 2.1), it is conceivable that it is harder to perceive as a focus marker than the pitch expansions in other sentence locations. In general, there are no known phonetic effects of focus on syllables that precede the focused word. On the other hand, there is a strong effect on phonetic material following the focused word. In particular, the pitch contour will remain below the pitch that would be employed in a sentence without a focus. Moreover, while word-internal stress is not erased, it gets dampened, resulting in a smoother pitch contour. In other words, the absolute post-focal pitch is lowered overall and its range is compressed. This form of compression has been dubbed post-focus compression (PFC). Xu et al. (2012) have shown that PFC and post-focus pitch lowering are
crucial for focus perception\(^1\). They compared native Mandarin speakers with native Taiwanese speakers as well as bilinguals of those languages. Mandarin is a language that does employ PFC while Taiwanese does not. Results of a perception study showed that Mandarin speakers are much better at detecting foci than are Taiwanese speakers. This is likely due to the additional focus coding achieved through PFC. In particular, it could be shown that Mandarin speakers drop to the accuracy level of Taiwanese speakers, once the focus is placed at the end of the sentence. In this case, no phonetic material follows and hence PFC cannot become operational. Interestingly, a final focus was perceived as neutral intonation in more than 25% of the trials, underlining the similarity of neutral intonation with the final focus contour.

Other than pitch, PFC also affected intensity but not duration in the study of Xu et al. (2012). This is in line with findings of Cooper et al. (1985) and Eady and Cooper (1986). The likely reason is that words that occur towards the end of a sentence tend to be shortened anyway and hence further shortening might render them unintelligible. In any case, these results suggest a tri-zone partition of focus on pitch (Xu, 2005, p. 235). In the pre-focus region, no significant changes take place. On the focused word, the pitch range gets expanded, modulating the slope but not the directionality of the word-internal stress. In the post-focus region, the pitch drops below the level it would assume if no focus was present (and therefore below its pre-focus level) and the pitch range becomes compressed. In parallel, intensity also gets lowered post-focally. Duration is not affected by PFC but it does extend significantly on the focused syllable (Eady and Cooper, 1986; Xu, 2005). The durational lengthening of focused words is not unexpected because speakers take time to perform the pitch range expansion (see Xu and Sun, 2002). The articulators (in particular the larynx) take more time to move in order to produce the greater pitch excursion, resulting in a lengthening of the syllable whose pitch range is expanded. Likewise, more time is required to lower the pitch onto PFC level after the focus. Therefore there exists a high correlation between the pitch and duration variation on a focused word. Again, there are no pre-focus effects on duration (Eady and Cooper, 1986). This suggests that duration is in fact a very salient on-focus cue. Interestingly, durational lengthening in conjunction with PFC seems to be a very salient cue for focus even if no significant pitch range expansion takes place. As Eady and Cooper (1986) report, there are hardly any significant pitch range expansions on the first word of a sentence it is focused. This is because English utterances tend to start at a rather high pitch level which then declines as the utterance goes on. Put differently, English utterances start close to the ceiling of the pitch range that is appropriate for an utterance that is intended to be a statement\(^2\). Of course it would be physically possible to exceed this ceiling but than the utterance might be mistaken for a question (a point I will elaborate on in the next section). Hence, the speaker can hardly expand the pitch

\(^1\)In this section, when talking about statements, I will use PFC to refer to both post-focus lowering and PFC proper.

\(^2\)As pointed out earlier, pitch ranges vary between speakers and hence no general threshold value can be given for such a ceiling.
CHAPTER 2

2.1. FORM

range any further. However, the focus can still be perceived accurately based on the lengthening of the focused word and consequent post-focus compression. We have thus identified three major components of phonetic focus marking in English statements: duration, pitch and PFC as well as a minor component which is intensity.

Questions

As a start, let me point out that according to def 1 there should not be something like focus in questions. It does, however, exist (Liu and Xu, 2007). This is a general problem for most theories of focus as they usually ignore focus in questions. Since it is not yet clear which exact functions focus performs in questions, we will define it in terms of its phonetic properties for the time being. We will therefore view focus in questions as the rise of the pitch contour to the pitch register that is typical for questions (see Liu and Xu, 2007, for justification).

With this auxiliary definition of focus in questions at hand, it is important to notice that focus in questions behaves differently from focus in statements. This is mainly because questions, at least in American English, are encoded by a steep pitch rise at the end if no focus is present (Eady and Cooper, 1986; Liu and Xu, 2007). More generally, questions need to end at a high pitch level. In such a context a mechanism like post-focus lowering would be odd since it would force the speaker to first drop his voice height below its usual level only to finally up it above that level. Not only would this necessitate great articulatory effort. Most likely, it would also result in a very slow production of the final word since changes of pitch level take time due to the inertia of the articulators (Xu and Sun, 2002). As one would expect on these grounds, we do not observe post-focus pitch lowering in questions. On the contrary, once the pitch has been raised due to focus, it stays up high till the end of the question (Liu and Xu, 2007). This way, focus and questionhood are encoded in parallel. In order to illustrate other encoding mechanisms, let us turn to figure 2.2.

First of all observe that the pitch after the focus on job in the question does not only stay at a high level but in fact keeps rising. Hence there is a combinatorial effect of focus and question encoding. The typical question intonation is still realized as rising, despite of the foregoing focus. Furthermore, we see that the question and statement intonation start to diverge on job even in the final focus condition. This effect is significant (Eady and Cooper, 1986; Liu and Xu, 2007), indicating that questions are cued even before the final rise. In other words, speakers employ a larger array of question coding devices than just final rise. As pointed out above, there is no post-focus pitch drop in questions. However, there is still post-focus compression in the sense that the lexical stresses following the focus in a question are realized within a pitch range that is narrowed compared to pre-focus intonation (Liu and Xu, 2007).

In addition to that, (Liu and Xu, 2007, p.1192) claim that question intonation also affects the pitch target of lexical stresses. In particular, they claim

\footnote{The pitch target is simply the underlying pitch contour that the speaker intends to pro-}
that question intonation forces rising pitch targets upon the lexical stress of pre- and post-focus content words and that this stress pattern differs from the one observed in statements. It should be noted, however, that the effect they report is only marginally significant ($p < 0.05$). Moreover, they only took the median final velocity\footnote{The final velocity is the slope of the intonation curve 30 ms before the offset of the (stressed) syllable. It is computed the change in semitones per second ($\text{s}^{-1}$).} of each speaker, sentence and intonation type (statement vs. question) and then compared the average of the medians across speakers. Hence, they likely reduced the variance in their data. More variance, in turn, might have led to a non-significant difference. Moreover, they considered content words. The content words in their study were 6 proper names, 1 adjective, 1 common noun and 1 pronoun. It is conceivable that the type of content word that is chosen may also have an impact on the tonal variance. In this particular case, the overwhelming presence of proper names may or may not have played a crucial role. To appreciate this, note that \citet{ramscar2013} have shown that names are particularly unpredictable. Hence, they may be harder to produce resulting in slower pronunciation. I am not aware of any studies that investigate the interaction of names and tone production or the (potential) differential impact of noun classes on tone production. I therefore tentatively conclude that the evidence for the impact of questions on lexical stress need to be backed up by further research.

\footnote{The actual pronunciation varies with each utterance due to the inertia of the articulators and neuromuscular deficiencies. However, the underlying pitch target can be estimated by averaging over the tonal contours of different speakers who have produced a word repeatedly. This is what Liu and Xu have done.}

Figure 2.2: Difference between American English statement and question intonation. The focus was on either job or La Massage and the sentence was pronounced as either a statement or a question. The curves are averaged over 8 repitions of 5 subjects. Graphic taken from Liu and Xu (2007).
In sum, questions in general differ from statements in that they end with a high pitch. If a word within a question is focused, the pitch range is expanded and the duration of the focused word is lengthened (Eady and Cooper, 1986). In contrast to statements, the pitch does not drop after a focus in a question but instead stays high and even rises towards the end. PFC, however, is still functional. Pre-focally it seems that questions are distinguished from statements in that they expand the pitch range of content words. However, it remains to be clarified whether questions assign a rising pitch contour to all content words not under focus.

2.1.2 Acquisition

Recall that we assumed from the outset of this thesis that language is a form of human behaviour. Like most forms of behaviour it has to be acquired over time. An interesting question is what phonetic cues are indispensible for focus acquisition. This question also carries implications for phonetic focus marking in general as any answer to it will highlight the importance of the cues involved in acquisition. In this section, I will therefore review some important results about the acquisition of phonetically coded focus.

As pointed out in Section 2.1.1, there is a huge inter-speaker variability in the production of intonation, especially across the sexes. Therefore, infants whose mother tongue is a language like English, i.e. a language that encodes focus mainly prosodically, might struggle to learn the appropriate cues from their linguistic input. On the other hand, the prosodic input that infants get even in the pre-natal stadium is most likely very numerous and hence may guide the learners in their search for cues. In particular, it has been found that fetuses are able to perceive auditory signals at the age of roughly 24 weeks (Slater, 1998), implying that they are exposed to the tonal patterns of their native language very early on in life. Moreover, young children are able to discriminate between the rhythmic features of their native language and other languages and consistently produce phonetic focus features at the age about 3 years. This suggest that they acquire the rhythmic properties of their native language earlier than other features of that language. However, this still leaves the question of what exactly it is that children learn. In order to determine the features of the pitch contour that are crucial for the acquisition of tones and focus Gauthier et al. (2007) and Gauthier et al. (2009) conducted an array of neural network studies. They used self-organizing maps (SOMs) for their experiments. These were devised by Teuvo Kohonen, in order to model the specialization of neural clusters for similar input patterns (Kohonen, 1982). In particular, Kohonen based his model on the topological organization of the auditory cortex. Ever since their introduction, SOMs have been highly successful in modeling unsupervised acquisition of classes inherent in the input data. Because of their tight connection to the human brain architecture, they are very popular in biological sciences and still widely in use nowadays. In the following, I will briefly introduce the basic working mechanisms of SOMs, as this will facilitate understanding of the results obtained by Gauthier et al. (2007) and Gauthier et al. (2009). This exposition
is based on Kohonen (1982).

SOMs

The SOM is a neural network architecture. In neural networks, there are processing units (or simply nodes) that implement a certain kind of function. The best known amongst these functions are weighted summation and the sigmoid function. More generally, however, any mathematical function can be implemented in a node. The output computed by such a function is known as the node’s activation. The typical node in a neural network receives its (numeric) input from other nodes. Importantly, the connections between the nodes carry weights, meaning that not all nodes contribute equally to the activation of the receiving node. Moreover, every neural network has a dedicated number of nodes that directly represent the input values fed into the network from an external source. These special nodes are called the input layer. Let us call the weights leading to the node that is about to be activated $w_i$ for $i$ ranging from 1 to the number of input nodes connected to that node. Let furthermore $x_i$ be the activation of input node $i$. A node implementing weighted summation would be activated according to Formula 2.1.

\begin{equation}
\text{act} = \sum_i w_i x_i
\end{equation}

The activation of a node implementing the sigmoid function would be calculated extending up on this linear combination. Taking the value $\text{act}$ from equation 2.1 the sigmoid activation is given by Formula 2.2. The crucial difference between the two functions is that 2.1 is linear while 2.2 is not\(^5\).

\begin{equation}
\text{sigmoid}(\text{act}) = \frac{1}{1 + e^{-\text{act}}}
\end{equation}

In the following, we will not further delve into the concrete implementation of the node activation. In fact, Kohonen himself found that the activation function plays only a minor role in SOMs compared to the topological organization of the network, in particular the connections. Therefore, I will concentrate on the structural properties of SOMs. Usually, the nodes of a SOM are organized in a grid as shown in Figure 2.3. Ideally, the grid will be partitioned according to categories inherent in the data after learning. Although the grid is oriented, this orientation does not matter in practice. Every time a new grid structure is learned, the connections to the input layer are set randomly, resulting in a possibly different location of the partitions with each run. However, instances that are assigned to a common class in one run should be assigned to a common class in consecutive runs. Hence, the same number of classes should emerge with each run. Since the number of classes is usually not predefined (although

\(^5\)Readers interested in the exact relationship between these two functions are referred to Jaeger (2008). In there, the sigmoid function is derived from the linear combination and the technical properties of both are explained in more detail.
they can be) and their topological location is not fixed, the map is considered to be topologically self-organizing.

An important feature that leads to this self-organization is lateral inhibition as well as lateral excitation. It has been observed that neurons in the brain are arranged in column-like arrays (Kohonen, 1982). Each array has a neuron or neuron cluster that forms its centre of activation. The arrays are laterally interconnected. If one of them becomes activated it excites the arrays in its near proximity, potentially leading to an activation in them, as well. On the other hand, neurons that are somewhat further away get sent inhibitory impulses, suppressing their activation. As a result, the area around an activated neuron cluster is likely activated itself (though to a lesser degree) whereas neurons further away are being suppressed. This spread of activation and inhibition is locally restricted and usually only affects a small subpart of the network. The pattern of excitation and inhibition is schematically depicted in Figure 2.4. The carving out of the partition then works as follows: for each node the potential activation is calculated according to function implemented in that node. The node with the highest activation\(^6\) does indeed get activated, while all other nodes stay calm. The weight of the activated node is then updated so as to move it even closer to the input vector. The same happens to nodes in the proximity to the activated node, although to a lesser degree. This process amounts to lateral excitation. Lateral inhibition is achieved by moving the weights of nodes that are further away in the direction opposite of the input vector. As a result the region around the activated node has come closer to the particular input. If on a future iteration a similar input is presented to the network, a node from that region is likely to respond again, thereby reinforcing the bonds between the region and a particular type of input. If enough similar input instances are presented to the network, the topological structure will likely be partitioned as described above. Crucially, if there is only little or no internal structure in the input, the map will stay unorganized. Therefore, SOMs are a very reliable tool for unsupervised cluster detection in the input. Moreover, they are modeled very closely to the human brain architecture.

Simulation results

After this rather technical digression we can now turn to the results that are of interest for our understanding of focus. As pointed out earlier, one of the greatest problems that learners face when acquiring prosody is the high degree of variation in the pitch patterns they have to learn from. This becomes the most apparent when one looks at between-gender variation. Figure 2.5 shows focus patterns produced by a male and female speaker of Mandarin Chinese. They produced focus in two different locations of a sentence. The male frequencies

\(^6\)Alternatively, one may also choose a regression strategy where the input vector activates the node whose weight vector is closest to the input. Since inputs and weights are usually represented as vectors, closeness is precisely defined as the euclidean distance between the input and weight vector in \(n\)-dimensional space, where \(n\) is the number of input features. This regression strategy is what Gauthier et al. (2007, 2009) used for their studies.
2.1. FORM

Figure 2.3: Schematic grid structure of the SOM. Graphic taken from Kohonen (1982).

Figure 2.4: Schematic representation of excitation and inhibition in a SOM. The red dot in the middle represent the activated unit. This unit spreads activation in its vicinity (yellow area). It also sends out inhibitory signals to units further away (blue area). This is a local process and only part of the SOM is affected by it. The signals sent out by the activated unit spread in square-like fashion due to the grid structure of the map.
consistently lie below those produced by the female. This indicates that absolute pitch height alone would be a poor source of learning the prosodic marking of focus. However, when one takes the derivatives of the pitch contours, the similarities between the pitch patterns produced by men and women become obvious. This is because the derivative of a function eliminates all constants contained in that function. This means that the intercept, i.e. the point where the function graph crosses the y-axis, vanishes. For the case at hand this implies that the differences between men and women (and in fact all individual differences between speakers) that are due to the speaker’s normal voice height are erased. While at first sight this may appear to be a waste of valuable information, it may actually help learners to better learn focus as well as tonal categories in languages that possess lexical tones.

In order to test this hypothesis for the acquisition of lexical tones, Gauthier et al. (2007) trained two SOMs each with pitch contours and with their derivatives as input. The input was again taken from native speakers of Mandarin, a language which has four lexical tones, namely high, low, fall and rise. The task of the SOM was to cluster these tones into coherent categories. The first network had to perform a rough clustering, using 100 categories. Afterwards the prototypical vector of each category (i.e. the weight vector of the unit responding to the contours in that category) was submitted to the second SOM which only identified 4 categories. This procedure is reasonable assuming an exemplar-theoretic model, according to which instances of events are stored and later abstracted over \(^7\). In the present case an event would correspond to an utterance containing one of the relevant tones.

The results of Gauthier et al. (2007) strongly favour a view under which learners make use of the derivatives of pitch contours instead of just taking

---

\(^7\)For a more thorough explanation of exemplar-based theories see for example Bybee (2006).
Figure 2.6: Simulation results for focus acquisition. The light grey bars represent the results obtained with derivatives of pitch contours as input whereas the dark grey bars show the results obtained from raw F0-contours. The first pair of bars labeled focus0 represents the results on pitch contours without any focus. Graphic taken from Gauthier et al. (2009).

fundamental frequency as the sole source of information. The authors found that the final network that was trained on F0-contours did not manage to come up with coherent categories at all. Hence this network could not detect the four tones in the input. The network trained on the derivatives, on the other hand, managed to perform this classification perfectly. These results are of course conditioned on the performance of the bigger network with 100 categories. The performance of this first network was above 90% when trained on derivatives whereas the network trained on raw pitch contours mostly achieved scores below 70%. The performance was assessed by measuring how many percent of the input tones were classified into a cluster of the same tonal category. With these results at hand, Gauthier et al. (2009) set out to test in how far focus learning would depend on the information from pitch contours and their derivatives respectively. This time they trained several SOMs with different degrees of variation in the input. Here, we will concentrate on their final, most variant experiment. The SOM contained 900 units and the input was obtained from four male and four female Mandarin speakers. The position of the focus was either sentence-initial, sentence-medial or sentence-final and contained within the same 5-word utterance. There also was a condition in which no focus was present. The task of the network was to discriminate between these 4 conditions. The results of the simulation are shown in Figure 2.6. The performance is measured the same way as in Gauthier et al. (2007).

While sentence-initial and sentence-medial focus can be detected fairly accu-

---

8 One may well argue that it is not shown that learners explicitly compute derivatives. However, some mechanism that manages to abstract away from differences in individual voice heights is clearly needed. The derivative is the most convenient mathematical tool at our disposal to model such a mechanism.
rately from the derivatives, the performance for the non-focus and sentence-final condition is poor. Raw F0 delivers acceptable results only for the sentence-initial position but consistently scores lower than the derivatives. These results may seem rather discouraging in comparison to those achieved for lexical tone learning. Note, however, that they do not come unexpected. First of all, it has been explained in Section 2.1.1 that even humans find it rather difficult to discriminate between no focus and sentence-final focus. This is mostly because both of them lack PFC, which is an important cue in languages that happen to employ it. For the simulations this means that fairly many units were ambiguous between no focus and final focus. Hence, any input vector to which such units responded would automatically count as misclassified. In other words: there were much fewer units that were able to recognize final focus, say, than for example sentence-initial focus. Moreover, we have seen that there are many more cues for focus that are different from pitch, such as duration and intensity. The SOMs in this simulation did not have any access to these information sources. Thus, the input they got was in fact impoverished. If all information had been made available, they might have performed better. However, the way the study was conducted is in fact optimal for our purposes as it allows us to compare focus acquisition to tone acquisition. It shows that duration and potentially other factors are crucial for the acquisition of focus. This is expected since duration is also a very important perceptual cue. Moreover, the studies by Gauthier et al. indicate that it may be imperative for the learner to first learn lexical tones and then acquire focus on top of that. This is because lexical tone seems to be relatively easy to learn and interacts with focus. In languages like Mandarin, it may influence the way the focus has to be realized since tone production has coarticulatory effects on focus production. Even in languages like English, that only employ lexical stress, focus expands the pitch range of the stressed syllable. Hence the stress pattern of the language and its metric properties have to be learned first in order to correctly make use of focus later on.

Finally, the fact that the derivatives of the pitch contours gave rise to better results throughout makes any theory hard to defend which claims that focus is a phenomenon solely based on extrema in the pitch contour. On the contrary, this finding supports the view that focus is a complex phenomenon that manifests itself continuously in the speech signal.

2.1.3 Syntax

Apart from phonetic focus marking, all languages that have been investigated so far also show some kind of morphosyntactic focus marking, though this may be optional. Moreover, there are also languages which solely employ morphosyntactic coding devices to mark focus, such as the African languages Hausa, Gürünțüm and the South-East Asian Malay (see Hartmann and Zimmermann, 2009, and references therein). For example, Hartmann and Zimmermann (2009) report that in Gürünțüm the morpheme \(a\) is used to mark focus. Their analysis should be taken with a grain of salt, however. First of all, they never really explain what they take focus to be and just assume their example sentences in
which a occurs to be focused. In particular, they assume that focus occurs in isolated statements without a foregoing question or assertion that might trigger the focus (Hartmann and Zimmermann, 2009, p.1362). This is at odds with what has been reported in the experiments cited in Section 2.1.1. Moreover, the morpheme a never occurs in future tense. The authors stipulate that focus may only be placed on events that are already concluded by the time of speaking. Of course, this is a somewhat ad hoc explanation. Leaving these issues aside, based on the literature cited in Hartmann and Zimmermann (2009) we may still tentatively assume that there are languages that use affixes to mark focus at least on some occasions.

For our current discussion, which is centered around English, what is more important are certain syntactic constructions that have long been thought of as marking focus. The main construction which has repeatedly been brought into connection with focus are it-clefts (see Huber, 2006). An example is given in (1).

(1) a. It was the dog that chased the cat.
   b. It was the cat that the dog chased.

The basic structure of a cleft is that it begins with It be+inflection and is then followed by a noun with a relative clause. It-clefts are seen as focus markers for two reasons. First of all it has repeatedly been claimed that the noun which precedes the relative clause needs be phonetically marked as focused (by the mechanisms explained in Section 2.1.1). While this may intuitively hold in most cases, I am not aware of any studies that would reliably have confirmed this claim 9. In principle, there is nothing wrong with the idea that focus can be encoded by phonetic and morpho-syntactic means at the same time. After all, phonetic focus marking also involves a range of redundant coding mechanisms, as described in Secion 2.1.1. In fact, due to the noisy nature of our environment, but also because listeners need as many evidence as possible to discriminate between different linguistic behaviours, it is indeed very beneficial to use redundant encoding. I will further elaborate on this issue in Sections 3.1 and 3.3; but see also Xu (2005).

Second, it has been been claimed that the it-cleft construction in itself may mark the noun it contains as being focused. This idea has been driven so far in generative syntax as to claim that there is a dedicated focus position in speakers’ language production faculty (Rizzi, 1997), a proposal that no one has ever attempted to confirm empirically 10. While the exact function of it-clefts is hard to determine, it may be insightful to note that most of the syntactic coding mechanisms that have been associated with focus place the focused word towards the edge of the sentence. As we have seen in example (1), this peripheral

---

9In general, one should be rather careful with claims about it-clefts. As Michael Ramscar pointed out to me they are hardly used in conversational speech. Due to their low frequency, speakers’ intuitions may be misguided as they may not encounter representative examples of it-clefts often enough.

10On the contrary, Abels (2012) has recently shown that even within the generative framework this proposal is rather absurd.
positioning mostly takes place at the left sentence edge, although sentence-final positioning has been reported as well. It would be interesting to investigate whether there are general sentence-peripheral properties that somehow make these locations suitable for focus marking or other functions. I will not pursue this general issue further here, but briefly return to it in Section 2.2.2. For now it suffices to note that there are certain morpho-syntactic mechanisms that likely encode focus (or at least fulfill a similar function).

An interesting idea in relation with focus and contrast has been put forward by Neeleman et al. (2009). They tease apart the notions of contrast and focus. This is interesting, since traditionally analyses of focus have brought it into an intimate connection with contrast (see e.g. Rooth, 1992). The idea of Neeleman et al. (2009) is that focus and contrast are communicative functions (or grammatical ones, in their terms) that exist independently of each other but may interact and occur simultaneously. This view might indeed turn out to be a fruitful approach when trying to explain why certain syntactic configurations are better suited for marking focus than others. The idea would be that focus and contrast enforce each other to a certain extend and hence the one makes the other more likely to occur. While this idea is of course very sketchy and would need to be solidified, it readily applies to the case of it-clefts. After all, this construction is usually assumed to delineate the noun it contains from another, mostly explicitly mentioned, alternative noun.

2.2 Function

2.2.1 Views from Semantics

Within the field of semantics, focus is standardly viewed as inducing a set of alternatives. More precisely, the alternatives are expressions that could in principle substitute the focused word\textsuperscript{11}. The idea of alternative semantics goes back at least to Hamblin (1973). Its application to focus has been most prominently pronounced in Rooth (1985) and Rooth (1992). Consider the sentence in (2-a). Let us assume that the word in uppercase be focused. Then the alternatives to that word would include, but not be exhausted by, the expressions in (2-b).

\begin{enumerate}
\item a. JOHN loves Mary.
\item b. Peter, Harry, Carl, Francis, Richard
\end{enumerate}

The concrete set of alternatives varies between conversational situations and is assumed to be restricted pragmatically. However, neither Rooth (1992) nor any later work explicate how this contextual restriction is supposed to take place. In fact, no method is known that would allow us to determine the impact

\textsuperscript{11}In very rigorous semantic terms alternatives are model-theoretic objects. Strictly speaking, the alternatives in example (2-b) should be individual constants. However, since names are more convenient for illustration than $c_1, \ldots, c_n$, I will use alternative expressions in my examples throughout. The reader may substitute them with the appropriate model-theoretic terms if he so desires.
of conversational context on language processing and production (Bell et al., 2009). If we had such a method at hand, it would most likely mean a huge step forward in our understanding of language. However, until this point is reached, any theory relying on the help of conversational context is standing on wobbly legs.

There are other theories to focus semantics which either split every utterance in a foreground and a background (Krifka, 1991) or assume that focus introduces some kind of presupposition (Geurts and van der Sandt, 2004). However, they are very akin to Rooth's proposal in that they also invoke an abstraction over possible alternative expressions (or model-theoretic objects, if viewed on a strictly semantic level). All these theories suffer from many weaknesses. The most obvious one is that the term *alternatives* is not well-defined. While Rooth (1992) uses it to refer to equivalence classes of the same model-theoretic type, Roberts (2012) for example uses it to refer to propositions. Likewise, alternatives could also be alternative utterances and hence not have anything to do with model theory. Another problem is that alternatives have prominently been claimed to be a distinctive feature of focus and the function of focus has been reduced to “indicat[ing] the presence of alternatives that are relevant for the interpretation of [the focused expression]” (Krifka, 2008). This is very problematic even within the framework of model-theoretic semantics. Alternatives are assumed to play a crucial role for implicatures, as well. This means that they cannot be distinctive for focus. Moreover, under a more communication-centred view like the one to be presented in chapter 3, alternatives are so vital for language in general that it would be impossible to tie them to any particular phenomenon. Hence, even if one wishes to account for focus using semantic alternatives, they are certainly not discriminative for focus. Moreover, one would have to find a consistent notion of alternatives.

Leaving the realm of semantics proper, the idea that focus invokes alternatives does not conform with experimental findings. It is known that semantically, but also phonetically related words *prime* each other. Priming means that the presence of a certain word makes other words, that are related to it, more accessible for lexical retrieval. If focus did induce alternatives, one would expect that this priming effect gets enhanced once the prime (i.e. the word that appears first) is focused. However, this prediction is not borne out in experimental studies (see e.g. Gernsbacher and Jescheniak, 1995, and reference therein). The psychological plausibility of focus-related alternatives may therefore be called into question. Finally, one of the most severe problems of alternative-based accounts of focus is that they are centered around particular constructions that involve so-called *focus-sensitive particles* such as negation, *only* and (arguably) *even*. These are called focus-sensitive because it is assumed that (i) they always need to co-occur with a focus (Rooth, 1985) and (ii) their semantics is not fixed but depends on the focus they co-occur with. As for (i), there is no corpus study or experiment that would have tested this assumption. While it may intuitively make sense, it has to be viewed as a stipulation. To understand (ii) better let us take a look at (3) which is a slight modification of (2-a).
CHAPTER 2  2.2. FUNCTION

(3) Only JOHN loves Mary.

The semantics of Only would in this case depend on the alternatives listed in (2-b). In particular, sentence (3) would state that neither of the alternatives is in love with Mary but John is.

At this point it is crucial to notice that only in the presence of focus-sensitive particles do the above theories ever make use of the focus alternatives. In all other situations the alternatives serve no purpose. Hence, these theories do in fact not offer any explanation for the presence of focus in general. The discourse contribution of focus is usually described as highlighting certain expressions or making them more prominent. This is of course far from concrete and in fact not part of the formal theories. I hence conclude that these theories have a rather weak contribution to make to the discourse function of focus.

One notable exception in the realm of semantics is Schwarzschild (1999). Schwarzschild proposes to view focus as a mere marker of non-givenness. Givenness is a rather intricate technical notion in his work but can be summarized as previous mention of an expression or the fact that an expression is in some way entailed by the foregoing discourse. The account proposed by Schwarzschild (1999) is refreshing insofar as it breaks with conventional semantic wisdom. Arguably, since it does not concentrate too much on focus-sensitive particles, it is the most general account of focus in formal semantics. However, selling focus as a mere marker of non-givenness means selling it short as we will see in the following sections.

2.2.2 Views from Psycholinguistics

The investigation of focus in psycholinguistics and psychology has mostly concentrated on the effects that focus can have on memory for words as well as their phonological and semantic associates. Also, the question of whether and to what extent focus might attract the listener’s attention to the focused word has been a major topic of interest. It has been suggested that memory enhancement might come about by virtue of the increased attention that listener’s pay to the focused expression (Birch and Garnsey, 1995). The general interest in the connection between focus and memory arose because interlocutors need to store certain information about what has already been said. This is necessary in order to be able to later refer back to the already mentioned expression, for example in anaphora such as in (4).

(4) I met John yesterday. That was a real surprise. I hadn’t seem him in ages.

In this short narration the pronoun him needs to be linked to the previously mentioned expression John (or alternatively to the person that is described by the name John). This would obviously not be possible if the discourse participants did not in some way store the information that John has been used before. Moreover, some meta-information needs to be accessible as well. For example, the speaker needs to recall that John is a male individual in order
to correctly use the pronoun him. While in this short example these needs are rather apparent (to the extent that it may even seem superfluous to mention them explicitly) it is certainly the case that real-life discourses are much more complex and reference to previously mentioned expressions takes place over longer stretches than the one in example (4). Therefore the idea that focus can be used to enhance memory for a certain word or the thing, person or event it describes has far-reaching implications. After all, focus would be a real discourse-structuring device as advocated by the founding fathers of information structure theory (e.g. Chafe, 1976).

Speech

I will first summarize what is known about the effects of prosodic focus on memory for words. These results will briefly be connected to more general insights about the connection between speech production and information status of the words produced. A more detailed discussion of this connection, however, will be postponed until Section 4.1 where it will be embedded into a more general theory of language production and processing. One of the clearest results about the memory effects of prosodic focus has been delivered by Gernsbacher and Jescheniak (1995). They could show that focus enhances memory for words when marked prosodically. To test for memory retention they asked their subjects whether a given word had appeared in the narration they had just heard. The target word was presented visually and reaction time together with answer accuracy were measured. The presentation of the target word took place either after the focused word or after a longer narrative segment following the focused word. In either case was the target word detached from the foregoing narration by a “filler phrase” such as y’know. Whenever a second narrative segment followed the target word it also introduced a new referent. This was done in order to test whether any effects induced by focus would remain even after the introduction of a new referent. Furthermore, this way it was checked whether focus not only enhances the memorizability of the word that it marks but also suppresses that of non-focused words. The overall built-up of the experiment is schematized in Figure 2.7. In the experiment, the focusable words would occur either after the focused word or after a longer narrative segment following the focused word. In either case was the target word detached from the foregoing narration by a “filler phrase” such as y’know. Whenever a second narrative segment followed the target word it also introduced a new referent. This was done in order to test whether any effects induced by focus would remain even after the introduction of a new referent. Furthermore, this way it was checked whether focus not only enhances the memorizability of the word that it marks but also suppresses that of non-focused words. The overall built-up of the experiment is schematized in Figure 2.7. In the experiment, the focusable words would occur either after the focused word or after a longer narrative segment following the focused word. In either case was the target word detached from the foregoing narration by a “filler phrase” such as y’know. Whenever a second narrative segment followed the target word it also introduced a new referent. This was done in order to test whether any effects induced by focus would remain even after the introduction of a new referent. Furthermore, this way it was checked whether focus not only enhances the memorizability of the word that it marks but also suppresses that of non-focused words. The overall built-up of the experiment is schematized in Figure 2.7. In the experiment, the focusable words would occur either after the focused word or after a longer narrative segment following the focused word. The target would correspond to one of the two focusable words in the critical conditions. Gernsbacher and Jescheniak (1995) justify this decision stating that many experiments, including trials they conducted themselves, have found a decrease in reaction times when the target was identical with a focused word but not when it was only a semantic associate (in fact, this pattern even carries over to the written domain where syntactic focus marking is tested, but see the following section). Hence any memory effects attributable to focus prosody would most reliably be captured using identical foci and targets.

The first condition the authors tested was prompting the target word after focus 1 (hence the target word would be identical to the first focusable word). They found a significant increase in reaction times when the word was focused. Error rates were very small anyway, so that there was not much space for im-
Figure 2.7: Schematic representation of the experimental set-up of Gernsbacher and Jescheniak (1995). A narrative segment ends with an optionally focused word which is followed by a filler phrase (e.g. *y'know*). The target word was prompted after the filler phrase. If the target word appeared after the first narrative segment, the second segment (the lower part of the graphic) would not occur. Conversely, if the target word appeared after the second segment, there would be no prompt after the first segment. Target words after the second segment could be identical to either the first or the second focusable word.

The trend goes towards even smaller error rates when the word was focused, nonetheless. This shows that focus enhances accessibility and therefore likely also memory for words over a small stretch of time. Another experiment prompted the target, again identical to focus 1, after the second stretch of narration. This means that a new concept had been introduced by the time the prompt occurred. If the first focusable word was not focused, reaction times and error rates increased as compared to when the prompt occurred directly after focus 1. On the other hand, if the word was focused, no such increase occurred for reaction times and only a marginal upward effect for error rates could be attested. Hence, focus seems to enhance memory for words even over longer periods of time. Notice that this is in line with the predictions made by a theory that takes focus to be a device that helps to manage discourse content. In particular, focus could enhance memory for words and/or concepts that are needed later on in the conversation.

Another experiment by Gernsbacher and Jescheniak (1995) tested whether focus would “surpress” the activation of other non-focused concepts. To this end, they varied the focus on the second focusable word and left the first one unfocused throughout. The prompt after the second narrative stretch would then test for the first focusable word. They found that when focus 2 was indeed focused, the prompt would be responded to marginally slower. However, the error rate increased significantly. The authors conclude that this indicates a “surpression” of the activation of the first word due to the focusing of the second word. In a final experiment, the second focusable word was focused throughout and the focus on the first focusable word was varied. The prompt always occurred after the second narrative segment and was alternately identical to either focusable word. If it tested for the second such word, reaction times and error rates were significantly smaller than for the unfocused first word. This is in line with the findings of the first experiment, viz. that prompting a word immediately after it has been focused will lead to a decrease in reaction times. The interesting part of this final experiment was when both words were focused.
In that case, subjects would respond equally fast and accurate independently of which word the prompt was identical to. This is interesting since the prompt always occurred after the second stretch of narration. While re-inforcing the idea that focus facilitates retention in long-term memory, it also shows that multiple foci can be stored simultaneously and that they do not inhibit each other. Moreover, no recency effect seems to be present. This leads the authors to dismiss a competition hypothesis according to which focus captures large memory resources that are consequently not available for the storage of other words or concepts anymore. I would like to add that this result even indicates that there is no focus-related suppression effect in general. Recall that the suppression effect that Gernsbacher and Jescheniak claim to have found was only marginal in the first place. Moreover, they did not control for recency effects or other kinds of linear or structural effects. Even if a suppression effect did exist, the authors fall short of explaining why there is no mutual suppression between the words when both of them are focused.

In sum, Gernsbacher and Jescheniak (1995) have convincingly shown that prosodic focus enhances the accessibility of the word that it marks. This happens both over shorter and longer time spans and is hence likely attributable to a better retention in memory of words that are prosodically focus-marked.

One point about the experiment of Gernsbacher and Jescheniak (1995) that often goes unnoticed is that the way they conducted their experiments actually disfavoured their predicted outcomes. In order to keep the focused and unfocused conditions as similar as possible, they spliced the unfocused versions of the words into utterances which were originally focused. This means that focus cues that do not operate directly on the focused word itself, such as post-focus compression, were still present even in the unfocused conditions. Hence, focusing effects towards better memory retention may still have been present in the unfocused condition. In other words: If the experiments were flawed by the splicing procedure, they were flawed so as to reduce the desired effects. Therefore, the results that the authors obtained are just the more convincing.

Finally, there is some evidence delivered by Dahan et al. (2002) that focus may serve to attract attention. More precisely, the authors found that focus may be used to shift attention from one discourse entity to another. They conducted an eye-tracking study in which four moveable objects were presented on a computer screen. Two of those were named by words that had the same onset (e.g. candy and candle). Alternately, one of these two words was pronounced in a first instruction. A second instruction would then always include a focused version of exactly one of the words (e.g. candle). Since humans start processing words even before they have been fully heard (Tanenhaus et al., 1995), the participants would either look at the actually named object or its competitor (whose name had the same onset). Dahan et al. (2002) found a general preference to look at previously unmentioned objects. On the second word, focusing had the additional effect that the competitor would be looked at if (i) it hadn’t

---

12Here, long-term memory refers to a time span of about 20 seconds. However, since no decay in performance was observed after this span, the retention effect likely lasts even longer.
been mentioned before (in line with the general preference for unmentioned items) or if (ii) it had appeared in a non-prominent position at the end of the first sentence\(^{13}\). Likewise, the actually named item would be looked at more intensely if it had occurred sentence-finally in the first instruction. Hence, not only does focus direct attention towards unmentioned items but rather to items that have not been made prominent before. In how far this attention shifting functionality of focus can be generalized is a question of further scrutiny.

**Text**

Textual studies have often been employed for two reasons: first, they are comparatively easy and cheap to conduct and, second, they allow to test syntactic structures that have been associated with focus-marking in the absence of phonetic cues\(^ {14}\). The text book example of an alleged focusing structure are it-clefts (see Section 2.1.3). However, other constructions, in particular those starting with *there be*+inflection, have been scrutinized as well. Concurrently with Gernsbacher and Jescheniak, Birch and Garnsey (1995) investigated the effects of these syntactic focus markers. Before we turn to discussing their results, a word of caution is in place. Birch and Garnsey (1995) do treat these constructions as focus markers. At the same time, however, they point out that a word is affected by these constructions in that they are “indicating that it is new information and calling attention to it” (p.234). This has two implications: first of all, focus may be far more complex and serve many more purposes than usually assumed. This is indeed what I am going to argue for in the rest of this thesis; second, there may be effects which are by themselves unrelated to focus but are due to other properties of the syntactic construction, e.g. as word order. Such effects may be hard to discern from focus effects proper, an issue that I will return to below.

For their experiments Birch and Garnsey (1995) used a probe recognition paradigm in which both the context and the target (probe) word were presented visually. There was a 1 second break in between the context sentence and the target word. Both the context and the target were presented in speeded fashion (3ms and 1.75ms respectively). The targets would either not be related to any words mentioned in the context (baseline), or it would be phonologically or semantically related. In accordance with Gernsbacher and Jescheniak (1995) there were also targets which were identical with one of the previously mentioned words. For identical targets, the effect could be replicated that those were responded to faster when the corresponding prime was embedded in a focusing structure. This held when the probe immediately followed the context as well as when there was a filled delay\(^ {15}\) of more than 30 seconds.

---

\(^{13}\) The authors actually frame their explanation in terms of thematic roles. They assume the theme, which occurs sentence-medial, to be more prominent than the goal, which occurs sentence-final. However, they did not discern thematic from positional effects.

\(^{14}\) At least there are no overt phonetic cues. However, readers likely superimpose a prosodic or rhythmic structure on a text when reading it (Fodor, 2002).

\(^{15}\) A filled delay is a time span in which subjects are presented with unrelated material in order to distract them. In the present case the filled delay expanded the narrative from the
second finding lends further support to the view that focus enhances memory over longer time spans so as to make words or concepts available for later reference in discourse. Note, however, that identical targets were the only ones that required an affirmative answer on part of the subjects. All other targets (unrelated, phonologically and semantically related) needed to be rejected due to the task design. Since rejections generally take longer, the identical target condition cannot be readily compared to the other ones. This may cast doubt on the effects found in the identical target conditions as these were computed in comparison to the unrelated condition. However, an additional naming study confirmed these effects, hence making them credible. Also, the experiments were later replicated in Birch et al. (2000) and the results were confirmed. In the replication trials the filled delay was shortened to 10 seconds. However, instead of reading a story continuation, subjects had to count backwards. This likely involves additional and (in part) different processing faculties and distracts the subject from the preceding story. Interestingly, subjects could much better recall a focused than an unfocused word. Again, this supports the view of focus as a discourse management device.

In the case of phonologically related targets the results remained rather inconclusive. While phonologically related targets took longer to reject, this effect was seldom tied to the presence of focus. In the condition with the filled delay between context and target and in the naming task, there were marginal interactions between focus and relatedness. These interactions were negatively correlated with reaction times, i.e. targets that were phonologically related to focused words were harder to reject. This is interesting because there was no general effect of focus on response latencies. However, the interactions were marginal at best. Moreover, there were no significant differences between the error rates. Such differences had been observed for identical targets, though. Hence, while it may be possible that focus enhances the activation of phonologically related words, the evidence presented is rather weak. Finally, semantically related words were not rejected faster when focused. Hence, there was no effect of focus on semantic relatedness. Recall that this was the reason for Gernsbacher and Jescheniak (1995) to only use identical targets. To summarize in Birch and Garnsey’s own words, “the results show that focus enhances memory for words and, in some cases, for their phonological characteristics” (Birch and Garnsey, 1995, p.250).

While these reception-based studies show that focus mainly supports word-level memory and, contrary to what has been claimed e.g. by Rooth (1992), has no effect on the activation of semantic associates, there have also been studies checking the effect of syntactic focus structures on production. Here, I will review the studies by Birch et al. (2000) and Kaiser (2010). While the former used stories as contexts, the latter used little dialogues. Participants read a context and then had to write a continuation. The hypothesis was that words which occurred in it-clefts (see example (1)) should be referred to more often in the continuations, given that it-clefts are generally taken to be focusing structures.
Kaiser (2010) also conducted a baseline experiment, in which no syntactic marking was used. This was to check the subjects’ general preferences. She found that her participants had a strong tendency to refer back to the immediately preceding subject. This preference persisted in the actual experiments whenever the referring expression was a pronoun. Notably, this had nothing to do with focus (i.e. the use of it-clefts). Kaiser (2010) cites several other studies which report about such a subject preference. It is well known that a word is made prominent by virtue of being in the subject position, independently of being focused. Syntactic focus does not seem to add to this effect. However, if the referring expression used in the continuation was a full NP it would often refer back to an alternative NP. This may seem to support focus accounts built on alternatives. However, in Kaiser’s experiments the focused word was introduced contrastively as in (5), where the dots indicate the position at which the participants had to start their continuations.

(5) A: The maid scolded the bride!
B: No, that’s wrong! It was the secretary that scolded her . . .

Hence, there was only one alternative and it was explicitly mentioned (in the above example this would be the maid). The effect that focused subjects led to more references to this alternative was significant (however, only if the referring expression was a full NP). On the other hand, if an expression refers back to a a subject, it tends to be pronominalized about half of the time. Hence, the overall rate of occurrence of this pattern can be estimated to be rather low. Moreover, a third experiment showed that if the alternative was not contrasted with the focused word, the alternative was almost never referred back to. For an example of such a configuration see (6).

(6) A: The maid scolded the bride.
B: Yeah, that’s right. In fact, she also scolded the secretary.

It therefore remains highly contentious whether the effect on the alternative is indeed due to focus or rather to the explicit contrast. The latter view has been put forward in the field of syntax by Neeleman et al. (2009). Further experiments are needed to draw any conclusion about this matter.

In sum, Kaiser (2010) shows that there is a clear bias towards subject-reference and this bias is not affected by syntactic focus marking. Basically the same finding has been made by Birch et al. (2000) in their experiments. They also used a continuation study in which they found that participants made more references to a focused noun than to an unfocused one. However, they counted the absolute number of references (i.e. one continuation may contain multiple references) and also did not control for subjecthood. Moreover, about 40% of the continuations did not refer to the target noun at all, a fact which is not included in their analysis. This means that the results of that study are hardly interpretable. However, they also replicated the study of Birch and Garnsey (1995) for identical targets. This time they had three conditions: first, the prime would appear in a there be+inflection or it-cleft construction, second it would
be prefixed by an indefinite \((a/an)\) or third it would be moved out of the subject position. This third condition was reacted to significantly slower than the other two. Moreover, there was no difference between the first two conditions. This again casts doubt on how much of a focus marker \(\text{there be+inflection}\) of it-cleft constructions really are. As I remarked before, this issue is far from settled. While a focusing functions of these constructions cannot be excluded it seems that the highlighting they instill upon the words they mark is strongly linked to some notion of subjectionhood or recency. Notice further that in fact most of the so-called “focus constructions” appear on the left or right edge of a sentence or utterance. While the left edge is easily associated with subjectionhood (at least in subject-initial languages like English), the right edge also seems to have some prominence to it. In particular, Ramscar et al. (2010) have shown that colours are better learned by children if they appear after the noun they modify. While the authors offer an analysis based on the relative informativeness of the colour word, it is conceivable that this effect of learning has repercussions on the perception of the sentence position the colour word occurred in, i.e. the left edge. Clearly, this is purely speculative\(^16\) and there are several certainly are several factors influencing the perceived prominence and memorizability of the sentence periphery. In fact, standard recency accounts would offer an explanation for both sentence edges. Be this as it may, the point in case is that although certain syntactic structures may fulfill the same functions as focus (e.g. memory enhancement), the focus component of this effect is hard to isolate as there are many additional factors involved. Hence, care should be taken when talking about “focus constructions”.

Summarizing, psycholinguistic accounts of what functions focus might serve have consistently found that it enhances memory for the word it marks, especially in the case of phonetic focus marking. This is likely crucial for future discourse segments in which words or their referents need to be refered back to. This back-reference obviously needs to be understood by all interlocutors for the discourse to succeed. Contrary to popular semantic theories such as Rooth (1985, 1992), no memory enhancement has been found for semantic associates of the focused word. However, other possible functions of focus such as the attraction of attention have been brought up. These ideas are plausible but need to be confirmed experimentally.

\(^{16}\)It is even more speculative in face of the fact that Ramscar et al. (2010) mostly talk about the relative position of the adjective and the noun. Their study involved no explicit test of how close the adjective was to the end of the sentence.
Chapter 3

Language and Information

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; this is they refer or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

(Shannon, 1948, p.379; emphasis mine)

It is a common consensus that language is used for the transmission of information. What kind of information language transmits is a matter ongoing controversy, however. Consequently, the way that information transmitted by language is formalized in theoretical work mostly depends on the theory in question. No theory-independent consensus has yet been arrived at. In most logical theory of natural language meaning, information is framed in terms of information states. These states are essentially descriptions of a particular situation\(^1\). There is no strict separation between information and meaning. For the time being, I will use them interchangeably. The meaning of a linguistic expression is determined by the effect it may potentially have on an information state. As a simple example consider (1) which will erase the possibility that Paul doesn’t own a dog. It will hence make the state more specific with respect to all situations in which Paul owns a dog. More commonly, it has been conveyed that Paul is the owner of a dog and everyone who didn’t know this before will now know this (given that they believe the speaker).

\(^1\)Technically, states are tuples consisting of a set of basic entities, an interpretation function and an assignment function. Other functions may be added.
Whenever an utterance manages to manipulate an information state, the state is said to be updated. Popular theories of meaning that rely on updates are dynamic semantics (Gamut, 1991, Vol.2) and inquisitive semantics (Ciardelli et al., 2012).

All forms of update semantics suffer from three severe problems. The first is a technical one. For these theories to work information states need to be constructed. This means that if they were to be applied to real-word situations one would somehow need to construct a model of these situations. This is hard to do since even every-day situations are relatively complex if they need to be represented formally. Moreover, the meaning of words in terms of update potentials and the interaction of these need to be captured somehow. The underlying assumption of update semantics is therefore the presence of a knowledge-base (for the representation of the situation) and/or an ontology (for the representation of the update potentials). Knowledge-bases are widely available nowadays but they are constrained to very specific topics and hence they are not readily useable for natural language in general. Ontologies, on the other hand, are known to only cover a very sparse portion of a language’s lexicon. Hence, neither technology is applicable for even moderately scaled linguistic problems. As (Charniak, 1993) pointed out already 20 years ago, this implies that all such systems are doomed to fail that build on the assumption that knowledge-bases or ontologies be available to support them.

A second big issue, that in some sense relates to the first one, is the explanatory leverage that symbolic systems in general are able to give us on how communication may work. As pointed out by Ramscar (2010), it is virtually impossible to tie the (deterministic) “meaning” of one expression to that expression itself. This claim may at first glance seem strange to researchers working with symbolic systems. After all, they do specify meanings as concepts or propositions or update potentials etc. However, this begs the question of what these “meanings” actually are. Just giving them fancy names like the ones just listed does not make them anymore intelligible. Specifying them explicitly through listing all possible instantiations of a particular type of meaning is also not feasible, as I pointed out in the foregoing paragraph. Even if such a listing was possible, it would still not specify the underlying communicative principle that was employed in the compilation of the list. To see this, consider any domain-specific ontology, for example from the medical domain. It may be extremely helpful for doctors when writing recipes and looking for treatments. From a general communicative perspective, however, it is essentially arbitrary. At the same token, since due to ambiguity there need to be different meanings for one expression, it is hard to imagine how people would choose the situationally appropriate meaning in a principled way if they were operating with a fixed set of meanings.

The third problem that update semantics suffer from is the narrow notion of information or meaning they employ. While they may to some degree be appropriate for the modeling of the exchange of factual information, they have
no leverage at social, emotional, corteous or other kinds of meaning. Those are, however, arguably more important and more frequently exchanged than factual information. A prime example supporting this claim is (2), perhaps the most frequent dialogue in the history of English.

(2) A: Hey, how are you?  
   B: I'm fine, thanks.

No factual information has been exchanged here. In fact, B would most likely not have said that he is miserable even if he was. Nonetheless, it would be odd and a potentially offensive breach of social conventions to start a conversation with a person one has just run into in any other way. After all, humans, just like other social species, constantly re-negotiate their social rank and their relationships with each other. One of the most important funtions of language is therefore social re-assurance. While apes groom each other to enhance trust and social coherence, humans talk about all kinds of things. The conversational topics may be irrelevant from a purely factual perspective, either because they are already common knowledge or because no facts are talked about to begin with. Nevertheless, the mere act of talking to each other strengthens the bond between the interlocutors. Likewise, gossiping is used to delineate oneself from other’s and enforce the coherence in one’s own social group. This suggests that a more encompassing theory of information is called for.

The natural choice of an encompassing theory of information happens to be information theory (Shannon, 1948). It’s main appeal comes from the fact that it treats information as a purely quantative measure. The interpretation of that information is up to the recipient\(^2\) and the interpretation that he employs. This is intuitively very akin to speech which also manipulates the air pressure around the speaker. These variations in air pressure are then first perceived and afterwards interpreted by the receiver. The interpretation that the receiver gives to the incoming signal depends entirely on him, i.e. on his personality, his state of mind and (importantly) on the experiences he has made in life so far. In the following I will carefully discriminate between information, a quantity carried by the signal, and meaning, the interpretation assigned to the signal by the recipient. This interpretation does not need to be one-dimensional. On the contrary, humans are able to extract a variety of meanings from a speech signal. Besides the factual meaning that may be conveyed, humans are very precise at judging emotions and body size (Chuenwatnanaprani et al., 2008) as well as sexual attractiveness (Xu et al., 2013). It is crucial to note that all these types of meaning are encoded synchronously. This implies that the signal itself is not one-dimensional (Xu, 2005, see also). It is therefore likely that receivers are actually able to filter the incoming signal and concentrate their interpretation only on the aspects that are relevant to them. Recent research strongly suggests that a theory that views linguistics signals, be they acoustic or written, as information and subjects their interpretation to the people perceiving them is

\(^2\)I use the term recipient to refer to people receiving a (linguistic) signal. This is not to be confused with the term receiver to be introduced below.
likely very adequate. I will elaborate on this point in Section 3.2. First, however, I will introduce the most important notions and definitions of information theory that also underlie modern linguistic theories.

## 3.1 Basic Notions of Information Theory

The advent of information theory came with the paper of Shannon (1948). Nevertheless some ideas underlying it have been around before. In particular, the idea that no message is meaningful in its own right but entirely depends on the interpretation a receiving system assigns to it goes back at least to Schopenhauer (1859). Schopenhauer argued that we cannot be sure of the very existence of the world around us since it is only through our mind that we conceive of it. Hence the world might just as well be a hallucination of our mind.

The more modern information theory of Shannon (1948) is mostly technically inspired. It certainly accepts the existence of the world. Its main concern is rather the transmission of a message from one physical system to another. The signal encoding the message may be discrete or continuous. We will focus on the discrete case here, since the basic units of language are usually taken to be either words, sounds, letters or syllables, all of which can be seen as discrete units. The theory assumes a source which emits a message following a stochastic process. This essentially means that the basic units of the message (the alphabet) are sampled with replacement. Crucially, this sampling process follows some probability distribution. If this distribution is for example binomial, the units being $A$ and $B$, and more probability mass falls on $A$ then this means that $A$ is more likely to occur than $B$. This has very practical consequences. Assume you were betting on whether $A$ or $B$ will occur. If you always bet on $A$, this would be your most profitable choice on the long run. Likewise, a recipient who takes in a message generated by a source that emits only sequences of $A$s and $B$s will be expecting to see $A$s most of the times (given that he knows about the source’s bias towards $A$). Hence, he will be more surprised to see a $B$.

Following the source, the next stage in the communication process is the transmitter. This object serves as an encoder for the message. Written text is one form of encoding a linguistic message. The transmitter in this case would be the script and the paper it appears upon. The encoded message (also referred to as signal) then has to pass through a communication channel. This channel can potentially be noisy, meaning that the signal may be distorted. In the presence of a noisy channel redundancy in the signal is needed so as to make it more robust against distortion. The signal is taken in by a receiver. The receiver decodes the signal and tries to restore the original message. Sticking to our writing example, the receiver would be the human eye. Distortions in the signal may be printing mistakes or blurring of the ink. The path from the transmitter to the receiver is known as the channel. This channel is tailored to a particular set of codes. In particular, the channel is designed to conduct all and only the symbols that that transmitter may send. This implies that
the receiver will always be able to read these symbols. Whether he manages to restore the encoded message is an independent question. Finally, the message arrives at its destination. Only there is it interpreted. The interpretation, however, is irrelevant to the communication process. Even if no interpretation takes place and the recipient disposes of the message without even looking at it, the communicative process can still be considered successful if the message that has been restored by the receiver is the one that was originally sent. A schematic representation of this model is given in 3.1.

A particular feature of this model, which makes it arguably very well-suited for linguistic communication, is that it puts the sender and the receiver on par. Every instance of communication is designed to transmit a message from the sender to the receiver. In fact, the communication is successful only if the message sent can be restored. Therefore, there is no sender- or receiver-centredness in this model. If we apply it to language data we hence make no prior commitment to linguistic functions being tailored to either benefit the sender or the receiver. In fact, the general assumption made here is that linguistic functions are designed to be intelligible and useful for both the sender and the receiver.

Information-theoretic Quantities

Having outlined the basic model of communication in information theory we can now turn to introducing the concepts we are interested in for this thesis. Note that all these concepts are grounded in probability theory. In order to work with them empirically, one needs to estimate the underlying probability distribution of the events one is interested in. I will address this issue below. For now we will just assume that this distribution is known. The first concept that has received a lot of attention in linguistic research recently is surprisal (Levy, 2008; Piantadosi et al., 2011). Surprisal reflects how unexpected an event is given its context. The lower the probability of the event in that context, the higher its
Definition 2 (Surprisal). The surprisal of an event $x$ is given by $-\log(x|c_1 \ldots c_n)$ where $c_1 \ldots c_n$ are instantiations of random variables representing the context of $x$.

Note that the logarithm does not have a fixed base. This is because it is a monotonic function. Hence any change in the base would just alter the value of the logarithm by a constant factor. In information theory and computer science it is common two assume a logarithm of base 2. The unit that expressed by a negative logarithm of values ranging from 0 to 1 is the bit. Intuitively, one bit corresponds to the information conveyed by knowing the outcome of a binary choice. A standard example is any system that can assume one of two states such as an on-/off switch. Knowing which state the switch is in means gaining one bit of information. We will adopt the convention of using bits throughout. This means that surprisal will also be measured in bits. Moreover, while the definition of surprisal assumes the presence of contextual random variables that the probability of an event is conditioned on, the context may in fact be empty. Technically, one can just use one random variable that deterministically assumes one state that leaves the distribution over events that can be conditioned on it unchanged. For convenience’s sake however, I will use the notation $P(x)$ whenever an event $x$ is not conditioned on any non-trivial context.

A particularly useful property of suprisal and other measures involving logarithms is their non-linearity. In particular, the difference of two values $\log(x)$ and $\log(y)$ varies according to the ratio of $x$ and $y$ rather than their absolute difference. This is what makes the logarithm useful for the modeling of probabilities of words in natural language. Most words are so infrequent that their probability is very low ($< 0.01$). Viewed on the probability scale, differences between the probabilities of these infrequent words are hard to detect, since probability is (asymptotically) linear in frequency. On the log-scale, however, even small differences in probability or frequency may matter for low-frequency words. As an example consider probabilities $p_1 = 0.05$ and $q_1 = 0.5$ as well as $p_2 = 0.0005$ and $q_2 = 0.005$. While their ratio is obviously the same, their absolute difference is not. Hence, the difference between $p_2$ and $q_2$ does not come out very strong on the probability scale. On the log scale, on the other hand, both differences are equal.

\begin{align}
|\log(p_1) - \log(q_1)| &= |\log(0.05) - \log(0.5)| \approx 3.322 \\
|\log(p_2) - \log(q_2)| &= |\log(0.0005) - \log(0.005)| \approx 3.322
\end{align}

For these reasons it seems that logarithmic measures (or more generally measures that vary according to ratios rather than linearly) are preferable when working with linguistic data. This is also why surprisal has proven so useful in recent linguistic research.

---

3This of course assumes that both states are equally probable.
4For convenience I chose to report the absolute differences so as to avoid negative differences.
To exemplify this and the following definitions, it will be helpful to introduce a running example. Say there are 10 balls in an urn, 2 of which are red, 3 of which are blue and 5 of which are yellow. Let us assume we conduct an experiment in which we are interested in the colours of the balls. In particular, we want to know what colour the ball has that we draw in a single trial. Assuming that all balls are equally likely to be drawn, we can compute the surprisal of drawing each colour as follows: first we compute the probabilities with which each colour occurs. These probabilities are $P(red) = 2/10 = 0.2$, $P(blue) = 1/3$ and $P(yellow) = 5/10 = 0.5$. Next we compute the surprisal of each colour which yields $-\log(0.2) = 2.3219$ bits for red, $-\log(1/3) = 1.585$ bits for blue and $-\log(0.5) = 1$ bit for yellow. Hence we find that red is the most surprising colour.

Another very important information theoretic concept is that of entropy. The entropy of a random variable measures the uncertainty about the outcome of that random variable. Intuitively, if the random variable follows a flat probability distribution it is harder to predict which value it is going to assume than when it follows a skewed distribution. This is exactly what entropy captures.

**Definition 3 (Entropy).** The entropy of a random variable $X$ that can assume states $x_1 \ldots x_n$ with the corresponding probabilities $p_1 \ldots p_n$ is given by

$$H(X) = - \sum_{i=1}^{n} p_i \times \log(p_i).$$

Figure 3.2 shows the development of the entropy for a random variable with two outcomes. The entropy is maximal if both outcomes are equiprobable, i.e. if we are maximally uncertain about what the actual outcome will be. Returning to our running example we can now quantify the entropy of the urn experiment. We already know the surprisals of each of the colours. Hence the entropy can be computed straightforwardly. Let us use $U$ as the random variable over possible colours.

$$H(U) = -p(red) \times \log(p(red)) - p(blue) \times \log(p(blue))$$
$$- p(yellow) \times \log(p(yellow))$$
$$= 0.2 \times 2.3219 + 1/3 \times 1.585 + 0.5 \times 1 \approx 1.49$$

This means that the urn draw contains 1.49 bits of uncertainty. Viewed from another angle, one could say that observing one draw from the urn gains the observer 1.49 bits of information. This happens by virtue of the fact that the observer’s uncertainty about the outcome of that draw is entirely removed.

Entropy has some interesting properties. As shown in Figure 3.2, it gets maximized if all outcomes are equiprobable. This captures the intuition that an event is most unpredictable if all other events are just as likely. This property is true in general. Moreover, it facilitates the calculation of the maximum entropy that a random variable can have. In fact, the maximum entropy of a random variable is uniquely determined by the number of its outcomes. To see this, assume any discrete random variable $X$ with $n$ outcomes. If the probability
Chapter 3 3.1. Basic Notions of Information Theory

Figure 3.2: Development of the entropy of a random variable with two outcomes. The two outcomes are $p$ and $q = 1$. The x-axis shows the probability of $p$. The entropy is plotted as a function of that probability. The entropy is maximized if both outcomes are equiprobable.

Mass is uniformly distributed amongst these, each will have a probability of $1/n$. Then the entropy is given by

$$H(X) = -\sum_{i=1}^{n} p(x_i) \times \log(p(x_i)) = -n \times (1/n \times \log(1/n))$$

$$= -n \times (1/n \times (\log(1) - \log(n))) = -(n \times 1/n) \times \log(n) = \log(n)$$

This shows that maximum entropy of any uniformly distributed random variable is exactly the logarithm of the number of its outcomes. Since it also is the maximum entropy of any random variable, this implies that the entropy of any random variable is bounded from above by this value; $H(X) \leq \log(n)$ for any $X$ with $n$ outcomes.

The final concept we need to introduce is relative entropy. It constitutes a measure of how much information is lost when one approximates one probability distribution with another one. Generally, it is used to measure the information
lost when changing the probability mass over a random variable from a distribution \( P \) to another distribution \( Q \). Another way to think about is the following: it quantifies how much information or noise needs to be added to distribution \( Q \) in order to get back distribution \( P \). So, essentially, relative entropy captures the information that is contained in \( P \) but not in \( Q \). Yet another take is to think in terms of signals\(^5\). If a signal for message \( M \) is generated by a transmitter according to distribution \( P \), relative entropy measures how much information needs to be added to the signal for the same message generated according to distribution \( Q \) in order to make both signals identical.

**Definition 4** (Relative Entropy). Given a random variable \( X \) with outcomes \( x_1 \ldots x_n \) and two probability distributions \( P \) and \( Q \) over those, the relative entropy of \( Q \) with respect to \( P \) is given as

\[
D(Q \mid\!\mid P) = \sum_{i=1}^{n} Q(x_i) \times \log\left(\frac{Q(x_i)}{P(x_i)}\right).
\]

That is, relative entropy is the weighted sum of the divergences between the probabilities assigned to the possible events. Relative entropy is also known as **Kullback-Leibler divergence**. Let us again exemplify it with our urn example. Assume we substituted one yellow ball with a red ball. Then the new probabilities of seeing each colour would be \( Q(\text{red}) = \frac{1}{3}, Q(\text{blue}) = \frac{1}{3} \) and \( Q(\text{yellow}) = \frac{4}{10} = 0.4 \). We can compute the relative entropy between the colour distribution in our manipulated urn and the original one as follows:

\[
D(Q \mid\!\mid P) = \frac{1}{3} \times \log\left(\frac{1/3}{1/2}\right) + \frac{1}{3} \times \log\left(\frac{1/3}{1/3}\right) + 0.4 \times \log\left(\frac{0.4}{0.5}\right)
\]

\[
\approx 0.17 + 0 - 0.089 = 0.081
\]

This shows that the probability distribution over the content of the manipulated urn diverges from the probability distribution of the original urn by 0.081 bits. Alternatively one could say, that 0.081 bits are needed to encode the information of the original urn that is not captured by the manipulated urn.

**Codes**

As of now we have been agnostic as to the nature of the signal. Generally, since information is standardly measured in bits, signals can unifyingly be represented as sequences of bits, i.e. strings consisting entirely of 1s and 0s\(^6\). Each basic unit of the message (e.g. a word) can hence be assigned its own unique bit sequence. A set of bit sequences with a mapping from to units of the message is called a **code**. This, however, leads to a problem for the receiver that has to decode the message. It does not know where one sequence starts and where the other one ends. This is because the bit sequences may be prefixes

\(^5\)This conceptualization of relative entropy is arguably closest to the noisy channel model in Figure 3.1.

\(^6\)This technique is in fact ubiquitous in every-day appliances. When sound waves are digitized for music records, for example, the wave gets transformed into a bit-sequence. The bits encode positions in time and amplitude that the wave has to pass through. In decoding the wave is reconstructed from these positions.
of each other. Let us consider an example with three sequences, namely 11, 01 and 1101. By the time the receiver has read the substring 11, it does not know whether this is the full sequences 11 or just the beginning of the sequence 1101. Worse even, it does not help to wait for the rest of the string to disambiguate. If 1101 was read at some point, the receiver would still not know whether it just read the sequence 1101, or a combination of 11 and 01. Several techniques have been introduced to counter this problem. One possibility is to introduce special delimiting sequences. Another one is to reduplicate each symbol within each sequence. These techniques have to be designed for each code individually, though. A more robust and universal way of dealing with the ambiguity problem is to use self-delimiting codes. These are designed so that no ambiguity can arise (each segment delimits itself). A common variant of self-delimiting codes are prefix(free)-codes (Li and Vitányi, 2008). These codes are constructed in such a way that no code sequence is a prefix of another one. The easiest way to construct such a code is using a binary tree. A simple example is given in Figure 3.3. While prefix codes eliminate ambiguity in the signal they have two major weakpoints. First of all, they are vulnerable to noise. Even tiny distortions may alter the signal in such a way that the original message cannot be retrieved. Another problem is that the alphabet of the code tends to become very large. Both these properties are diametral to the purpose of linguistic communication. In fact, ambiguity and redundancy are to be found everywhere in natural language. This is necessary since we are living in a noisy environment in which the signal gets distorted all the time. Therefore, prefix codes are probably a bad model of how natural language expressions are structured. Note, however, that prefix codes are inherently deterministic (although the distribution of their segments need not be). Once the receiver has read a sequence of the code it can be absolutely sure about the identity of that sequence. Human parsers, on the other hand, likely make use of the probabilistic structure that comes with ambiguity. Nevertheless, independent of the nature of the code, it will always be better to assign shorter sequences to units of the message that are used frequently and longer sequences to units that are used rather infrequently\(^7\). The reason is that

\[^7\text{A common way to assign short encodings to frequent message units is Huffman coding. Here the message symbols are arranged as leaves of a binary tree with more frequent symbols occurring further up in the tree.}\]
this way the overall signal gets shortened and hence more information can be transmitted per unit of time.

Now that we have introduced some relevant information-theoretic terms and explicated that they are inspired by the communication with technical devices, the question comes up to what extend all this is relevant to linguistics. I have claimed at the beginning of this chapter that information theory can be fruitfully applied to linguistic research. However, the relationship between this engineering theory and natural language may not be immediately obvious. In order to establish this relationship and disperse at least some of the doubts that the reader may or may not carry, I will survey some empirical evidence to support my claims in the Section 3.2. Before that, however, we need to hold on the probability distributions we will be working with.

3.1.1 Estimation

Up to now we have assumed the probability distributions we have been working with to be known. Knowing the true probability distribution underlying a stochastic process is a very comfortable situation to be in. Unluckily, in empirical research it hardly ever occurs. The reason is that we can only observe the output generated by the stochastic process. In principle this output should follow the underlying distribution if we can make as many observations as we desire; however, we cannot do so. Both our time and the number of events that are actually observable are limited. Hence, we have to retreat to making an educated guess about what the underlying probability distribution might be. This guessing process is know as estimation. The hope is that our observed data are close enough to the “real” probability distribution to allow us to arrive at sensible models about the phenomenon in question.

A very common form of estimation is maximum likelihood estimation (MLE). To gain a better understanding of this procedure let us first introduce likelihood.

Definition 5 (Likelihood). The likelihood of a model $M$ with respect to a data set $D$ is given by $P(D|M) = \prod_{i=1}^{n} P(d_i|M)$ where $d_1 \ldots d_n \in D$ are observed events.

Notice that we make no commitment to the form of the model. Any model that allows us to compute the probability of the data conditioned on that model

---

8It actually is a matter of debate whether there is anything like an objectively real distribution underlying a stochastic process. Many Bayesian approaches to statistics assume exactly that. Any divergences of the observed distribution are treated as noise under this view. Alternatively, one can also assume that a data set is structured according to its own distribution and that any divergences between two data sets from the same domain are due to a non-overlap in information content. This is exactly what relative entropy (def. 4) measures. The hope during estimation then is that we sampled a data set that needs relatively little additional information to describe any other data set from the same domain. The latter view is obviously closely connected to information theory. Nevertheless, this debate is mostly about conceptual matters and has almost no practical impact on the estimation process itself. The reader is hence encouraged to take the view that he prefers. Our exhibition of estimation techniques will not be affected by either choice.
is fine. In particular, there is no a priori need for the model to be a statistical one (although these models usually fare best in empirical research). The maximum likelihood estimate is then given by the model that maximizes the likelihood of the data. This is of course a purely data-driven estimation procedure. While any estimation technique that yields the MLE model is guaranteed to perform best on the observed data, there is no guarantee that it will deliver acceptable results on new data. In fact, if the training data is not a representative sample of the domain, the MLE model will generally perform very poorly. Experience shows that adding more data to the training set usually helps to improve the model’s quality (Halevy et al., 2009).

Let us now turn to concrete estimation techniques that are standardly used in computational linguistics and psycholinguistics. The most basic one is the frequency estimate. It simply boils down to counting the occurrences of one item of interest in a data set. Since we are dealing with language in this thesis, the items of interest will mostly be words but may occasionally also be letters. Hence, for a frequency estimate over a data collection we simply count how often each word or letter occurs. It is straightforward to derive a probability estimate from the frequency estimate. Let \( \# \text{words} \) be the sum of all occurrences of all words in our data. Furthermore let \( \# w_i \) be the count of all occurrences of a particular word \( w_i \) (for example rank the words alphabetically and let \( i \) signify the position in that ranking). Then the probability estimate is given as follows:

\[
\hat{P}(w_i) = \frac{\# w_i}{\# \text{words}}
\]

Here, I follow the convention to mark estimates by putting a hat above the estimated probability function. The above example shows the estimation technique for words. The technique itself is not restricted to words, however. It can also be applied to letters or strings of consecutive letter or words. To do this, simply replace item of interest in equation 3.6.

With this basic technique at hand, we can now turn to a problem known as language modeling. Language models aim to assign probabilities to strings of words (or letters). These strings need not be full sentences nor need they make sense in the language under scrutiny. Ideally, however, the language model will assign high probabilities to strings that are likely to occur in the language the model has been trained on and low probabilities to unlikely strings. As it turns out, simple per-word probability estimates are not very helpful language models. The reason that they do not keep track of which words are likely to occur together. Therefore they will assign high probabilities to strings that contain many frequent words but are unlikely to ever occur in written or spoken English. An example of such a string is (3).

\[
(3) \quad \text{The the the the the the the the the the the}
\]

In Bayesian statistics, the likelihood also involves prior probability over models. In practice it may be hard to establish such priors, however. With a uniform prior distribution over models the naïve Bayes assumption would be arrived at according to which \( P(M|D) = P(D|M) \). This is of course identical to the formulation of likelihood given in def. 5.
By virtue of the being the most frequent word of English, this string will receive a relatively high probability under a simple per-word language model. This is undesirable and hence a language model that takes word co-occurrence probabilities into account is called for. The standard answer to this challenge are n-gram models. An n-gram is a substring of n words or letters. N-gram models standardly estimate the probability of a word given the n−1 preceding words (although n-gram models based on following words can be built as well). Let \( W_n \) be a string consisting of n words and let \( W_{n-1} \) be that same string with its last word stripped off. The n-gram estimate is then given by plugging these terms into equation 3.6.

\[
\hat{P}(w_i|w_{i-1} \ldots w_{i-(n-1)}) = \frac{\#W_n}{\#W_{n-1}}
\]

It follows from the equation that per-word probability estimates are a special case of n-gram estimates. They are known as unigrams. In practice, 2-grams (bigrams) up to 5-grams are used. The probability of a string \( s \) with \( m \) words is then calculated as follows:

\[
\hat{P}(s) = \prod_{i=1}^{m} \hat{P}(w_i|w_{i-1} \ldots w_{i-(n-1)})
\]

Special markers at the beginning of the string may be introduced as substitutes for the missing foregoing words.

Intuitively, one would like to have n-gram models with wide spans so as to capture more contextual influence. This turns out to be problematic, though. The reason is that due to large number of words in a language many longer strings occur infrequently so that they are either not encountered at all in the data sample one is working with or that they are so rare that their distribution cannot be reliably estimated. For letters the situation is different. Here, 10-grams can easily be used. In order to alleviate the sparsity problem encountered with n-grams over words, several ideas have been developed that broadly fall into two classes. For the sake of completeness I will briefly introduce them but not further elaborate on this matter as it is not central to our concerns. I do, however, provide pointers to the relevant literature.

On the one hand their are smoothing approaches. These basically reserve a fraction of the probability mass for unseen events. Effectively, they assign probability mass even to strings that they were not trained on. The induced distribution therefore avoids very extreme peaks and valleys and hence has a more smooth contour. A detailed comparison of different smoothing approaches can be found in Goodman (2001). The second type of approach are back-off models. If a given n-gram can not be found by the language model, it will simply retreat to the closest (n-1)-gram. Essentially, this means cutting off the first word of the unknown string in the hope that the smaller string has been encountered before. Traditionally, back-off models include some parameters that penalize for backing off from the n-gram in question. With the advent of big data samples this procedure seems to become superfluous (Brants et al.,
Instead, parameter-free back-off seems to be just as good. In fact, back-off models are nowadays the preferred option, at least for large data collections, since smoothing techniques become too computationally expensive at sample sizes of several billions of words.

This concludes our exposition of the estimation of probability distributions over text and in particular language modeling. It should not obscure the fact that there is a large array of alternative language modeling techniques, as well. Some language models, for example, are based on latent structures, such as syntactic trees. For a good overview of these see e.g. Manning and Schütze (1999).

3.2 Is This Even Real?

One of the earliest relationships between language and information theory has been established by Zipf (1936). Zipf himself, of course, could not be aware of this relationship as he formulated his insights before information theory was even conceived of. However, he made the important observation that frequently occurring words tend to be shorter than infrequently occurring words. He ascribed this fact to a principle of least effort, according to which the effort in speaking (and other forms of behaviour) tends to minimized. Making frequent words short invokes less effort as it shortens the overall linguistic output. In information-theoretic terms that means that the code is optimized so as to transmit more information per time unit.

At this point I should make clear that I am not advocating information-theoretic considerations as the sole reason for linguistic behaviour. In fact, all the evidence we have gathered throughout this thesis (plus the one to come) suggests that language is an extremely complex phenomenon governed by many factors. For example, it is conceivable that the correlation between word length and frequency comes about through an interplay of code optimization and effort reduction (and possibly other factors). While I will mostly concentrate on information theory, I explicitly advise that other factors should always be tested as well. However, as we will see below, information-theoretic measures tend to have a high predictive power for certain forms of linguistic behaviour. At times these measures even remove the significance of other factors (see e.g. Baayen, 2010; Jaeger, 2010).

Coming back to Zipf’s observation, there has been some work extending upon it recently. Piantadosi et al. (2011) scrutinized the relationship between frequency and length in a cross-linguistic study using large corpora. There were 11 languages involved and for each of them the 2500 most frequent words were investigated. The author’s distinguish between frequency and information content. In order to have both on the same scale, they transform frequency into bits using the formula for surprisal (def. 2) with an empty context. This means that no conditioning took place. The probabilities were estimated using maximum likelihood estimates (see Section 3.1.1). The information content was assessed again using surprisal but with the context of one to three foregoing words. The
probability of the words given the contexts was estimated using n-gram-models (also introduced in Section 3.1.1). It was found for most languages and most contexts that information content showed a higher correlation with word length than mere frequency\textsuperscript{10}. Moreover, when the experiment was replicated for English using the BNC, the effect of information content was partialed out\textsuperscript{11} the values for of the frequency scores and vice versa. This revealed that frequency had no significant effect on its own when information content was taken into account. However, even with frequency already taken into consideration, information content still was a reliable predictor of word length. Piantadosi et al. (2011) furthermore found that frequency performs especially badly as a predictor for the length of words with relatively low frequency. In effect, low frequency words are almost treated on par by frequency as witnessed in Figure 3.4. This behaviour may be part of the reason why frequency alone is often a worse predictor than information content (i.e. contextually conditioned surprisal).

The study by Piantadosi et al. (2011) and the insight by Zipf (1936) suggest that language is structured so as to optimize the amount of information that can be transmitted per word. The fact that this seems to be true cross-linguistically indicates that efficiency is a trait of human communication in general. No innate

\textsuperscript{10}It should be pointed out of that all correlations were below 0.5 still, indicating that other factors are at play here, as suggested above.

\textsuperscript{11}Partialing out the effect of measure X of another measure Y can for example be achieved by regressing Y on X and replacing replacing Y with the residuals obtained this way. As a result the effects of Y are captured by X as far as possible and only the effects genuine to Y remain in the residuals.
constraints need to be postulated here. On the contrary, humans are likely very sensitive to the distributional properties of words (and other perceptual events, for that matter\textsuperscript{12}) and adjust their behaviour so as to most efficiently deal with these properties. This is supported by findings of Bell et al. (2009) and Gahl (2008) regarding the phonetic reduction of words that are expected in their respective contexts. First of all, the frequency distribution of English words is bimodal (Bell et al., 2009), given that one splits it between function and content words. This has to do with the distribution of open and closed class words. The distribution over each class has its own peak with closed class words (a.k.a. function words, such as the, of and and) being generally more frequent. Hence, both classes are affected differently by predictability and word frequency. In particular, word frequency by itself does not seem to affect the duration\textsuperscript{13} of function words. These words are the most frequent ones anyway and reside on top of the overall distribution. A binary discrimination between very frequent and rather less frequent function words turned out to be a significant predictor of duration. This indicates that while there is no general effect of word frequency on function words, there may be length differences between words of that are on the opposite ends of the distribution over this class of words. This is in line what has been reported in Piantadosi et al. (2011).

For content words (or open class words) the situation is very different. Their duration heavily depends on their frequency. Generally, more frequent words tend to be pronounced shorter. Bigram estimates with respect to the preceding word did not predict the duration of a content words. Note, however, that this does not mean that content words are immune to any kind of preceding context. Firstly, both (Bell et al., 2009) and Gahl (2008) used the Switchboard corpus to gather the materials for their studies. While Switchboard with its roughly 2.4 million words is one of the larger collections of spoken data, it is likely still undersized for the estimation of reliable language models. Hence, the missing effect of bigram probabilities may be due to the poor empirical fit of these probabilities. Moreover, we have seen that surprisal is often a more adequate measure of contextual predictability as it is log-transformed. (Bell et al., 2009), however, only used the raw bigram probabilities, which may have further decreased the predictive power of their estimates\textsuperscript{14}. Finally, Baayen (2010) showed that word frequency is likely only an epiphenomenon of several

\begin{footnotesize}
\textsuperscript{12}Berns et al. (2001) could show that the area of the human brain usually associated with the detection of irregularities, the ventral striatum, reacts heavily to only small manipulations in the taste of a fluid that their participants received in regular intervals. In particular, they would react whenever the less expected (more infrequent) taste was presented to them. Berns (2005) presents many more examples in which humans engage more strongly with unexpected events than with ordinary ones. Hence, the effort that they put into dealing with unexpected events (and the neural activity associated with this effort) are greater than the one invested into treatment of rather expected events.

\textsuperscript{13}Bell et al. (2009) estimated the average duration of a word from the average duration of the phoneme it contains. This way they could circumvent sparsity issues during estimation. Moreover, the phone length was thus estimated over different contexts. On the downside this may mean that idiosyncrasies of certain phonetic contexts, that are for example due to coarticulation, may not be captured in their study.

\textsuperscript{14}Incidentally, the authors did use a log-transformation for word frequency.
\end{footnotesize}
contextual predictors to the left of the word in question that conspire to make a frequent word appear more often. In order to do so, he used response latencies from a lexical decision study since such latencies are known to be rather strongly correlated with a word’s frequency. Once the contextual effects are partialed out, only little predictive power remains for frequency. Baayen (2010) goes on to show that morphosyntactic context is better suited for learning to predict lexical decision latencies than learning algorithms that learn from word lists (i.e., mere frequency of occurrence). Hence, while the results of Bell et al. (2009) show a strong effect of word frequency on the duration of content words, they do not preclude the presence of contextual effects (nor do they show the presence of such effects—more fine-grained investigations are needed here).

Further effects that (Bell et al., 2009) found for content words were a strong effects of the bigram probability with respect to the following word, an interaction of this effect with word frequency and an effect of word repetition. After repetitions, content words were generally shorter, although this effect levels off after the first repetition. Notice that repetition likely enhances the predictability of a word since that very same word has just been mentioned and should therefore still be highly active in the interlocutors’ word retrieval facilities (I will stay agnostic with regards to the exact nature of these).

In the case of function words, there was a significant effect of both bigram probabilities with respect to the foregoing and the following word. The effect for the previous word had a strong interaction with the binary frequency discrimination, declining for relatively infrequent function words (a similar, but much weaker interaction was found for the following word, as well). This indicates that the effect is largely due to collocations. Certain very frequent constructions such as kind of make the function word involved in them highly predictable.

There was no repetition effect on function words, likely because due to their general high predictability they tend to be very short and reduced to begin with (see also fn. 16 of this chapter).

In another paper, Gahl (2008) tested the effect of lemma frequency on homophone duration. The idea is that while homophones sound alike and should hence not be distinguished if interlocutors only pay attention to words as phonetic or syntactic entities, homophones can be discriminated based on their lemmata. An important observation of Gahl (2008) is that most homophone pairings are between high- and low-frequency words. As we know from Piantadosi et al. (2011) and Baayen (2010), frequency strongly correlates with contextual predictability (crucially, however, contextual predictability is even more

---

15 This does not necessarily mean that humans use some kind of right to left online processing. The significant effect of reversed bigrams can also be explained by the very nature of the model itself. It introduces many independence assumptions and some of those may be relaxed through reversed bigram estimation. Hence, no conclusion based on this effect can be drawn with respect to processing directions. I thank Khalil Sima’an for pointing this out to me.

16 At this point it may instructive to remind the reader of the findings by Xu and Sun (2002) that speakers are limited in their maximum articulation speed and that their usual speech tempo is fairly close to that maximum speed. Hence, an inverse linear relation between number of repetitions and a word’s duration should not be expected. Physiology (in particular the inertia of the articulators) precludes this.
informative about a particular word). Hence, homophone pairs optimize the linguistic coding system so as to avoid the use of very long word forms and by the same time make the homophones separable based on the very different contexts they are likely to occur in. In her data, Gahl (2008) not only found bigram-effects and similar to those of (Bell et al., 2009) but she also again confirmed that the noun-ratio of each word. The relative amount of nouns compared to other POS classes that instances of a particular orthographic representation encode, tends to increases duration. Furthermore, she could show that homophones close to prosodic breaks were rather long. This effect could not be detected by Bell et al. (2009) as they explicitly only considered sentence-medial words. Finally, the important finding of Gahl (2008) is that lemma frequency of a homophone indeed reduces the duration of that homophone. This finding should be taken with a grain of salt, however. First of all, the models fitted by Gahl (2008) are rather weak (\(R^2 < .5\)). Secondly, it bears reiterating that frequency effects are highly collinear with contextual predictability effects. In the case at hand this means, that not necessarily the mere lemma frequency but rather the different contexts that the homophones occur in (and which make them more or less predictable) caused the divergence in duration. Be this as it may, the contribution of this work should not be underestimated as it points out that semantic predictability is just as important as syntactic predictability. The same point has been made even more convincingly by Ramscar et al. (2012). They used nonce words that where phonetically similar to existing words. Of these existing words, one followed a regular and one followed an irregular inflectional paradigm\(^{17}\). They then presented their subjects with stories that enforced the semantics of one of the existing words. However, the verb occuring in the stories was the nonce word. At the end, participants were presented with a sentence that reflected on the story using the nonce verb in one of the two inflectional forms. Their task was to verify whether the sentence correctly reflected on the events described in the story. Notice that in this case, it really is semantic (and phonological) association that causes the effect, since the subjects could not possibly have had any prior experience with the nonce verb. Taken together, all these studies show that using frequency and contextual predictability (i.e. suprisal), many linguistic behaviours can be explained that would be hard to render in terms of grammar or other rule-based systems.

Leaving the word level and turning to reduction in syntactic contexts, it has been found by Jaeger (2010) that syntactic reduction often occurs when a syntactic item is redundant from an information-theoretic point of view. In particular, certain words may occur in order to spread the information density of the utterance or they may not occur in order to condense the information density. An example is given in Figure 3.5. The underlying idea is that lan-

\(^{17}\)In the light of the findings of Ramscar et al. (2012) and other evidence gathered from predictability and code form, one may indeed wonder how sensible it is to talk about regular and irregular inflectional paradigms. After all, even irregular word forms follow a pattern that is by now means arbitrary.
Language is optimized so as to transmit as much information as possible without at the same time exceeding the channel capacity. Moreover, information density should stay constant throughout, so as to ensure smooth signal encoding and decoding. In linguistic communication it is obviously hard to define the channel capacity since no physical channel is present. However, viewed from a surprisal perspective the channel capacity may be construed as the maximum amount of information that the interlocutors can steadily exchange without threatening the success of the communicative process and hence their dialogue. In fact, Jaeger (2010) measures information as surprisal. He used this measure to predict the occurrence of the complementizer *that* before embedded clauses. Using logit mixed effect models as described in Jaeger (2008), he could show that information density is a significant predictor for the occurrence of *that*. Specifically, it turned out that the higher the information density at the beginning of an embedded clause (i.e. the more unpredictable the onset of that clause) was the more likely was *that* to occur. Other significant predictors included the realization of the matrix subject (full NP vs. pro-form), length of the embedded clause, frequency of the matrix verb and preceding pauses (amongst others). That frequent matrix verbs reduced the probability of *that* should come as no surprise at this point of our discussion. After all, the subcategorization preferences of verbs that occur with high frequency should be well-known to competent language users and hence the occurrence of an embedded clause following such a verb is rather predictable. What is interesting then is that pauses preceding the embedded clause increased the probability of *that*. In effect, pausing might contribute to the effect that Jaeger (2010) ascribes to *that*; namely, it might spread out information density and hence make it easier for (i) the speaker to access the words he intends to use in the embedded clause and (ii) for the listener to process the utterance.

Apart from the effects reported here, Jaeger (2010) found other effects from several control variables that he used to strengthen his results but also to test other theories. He points out that while information density emerged as the best predictor, many others are also at play. Hence, instead of advocating one theory as the true one it is likely that many theories are right to some extent. However, it seems that information-theoretically grounded approaches are superior if theories are looked at in isolation.

Another investigation of surprisal effects in syntactic constructions was conducted by Levy (2008). These basically solidify what we have explored so far by applying surprisal as a predictor for perfective verb occurrence in a language with verb-final embedded clauses (German) and to syntactic ambiguity (garden-pathing) as well as anaphora resolution (in particular personal pronouns). I will not review these findings in detail here, as I think that the reader should by now be fairly convinced that surprisal can indeed be applied to each of these phenomena. Also, reporting the results would force me to introduce another estimation method, which would exceed the scope of this thesis\(^\text{18}\). What I would

\(^{18}\text{Levy (2008) uses PCFGs for the estimation of surprisals, hence the context he employs is slightly different from what we have seen so far. The interested reader is referred to Charniak}\)
Figure 3.5: Plots showing the information (surprisal) per word in a sentence. A channel capacity was stipulated and is indicated by the solid grey line. The upper plot shows a sentence in which the channel capacity would be exceeded if the complementizer was left out (red line). On the other hand, if the complementizer is used, the information transmitted stays fairly uniform throughout the sentence. The plot at the bottom shows a sentence in which the introduction of the complementizer would lead to a drastic undershoot in information content. Graphics taken from Jaeger (2010).
like to focus on instead is a rather remarkable theoretical result delivered by Levy (2008). In many theories of language processing, some latent structure, such as trees, is assumed. Since sentences are processed incrementally (Tanenhaus et al., 1995), the distribution over these latent structure will change with every new word. As long as no input has been processed, all structures are in principle possible (although there is no need to assume an a priori uniform distribution). With every new word the distribution over the same structures gets updated, possibly assigning no probability mass at all to structures that can be ruled out at this point. This means that the size of the transition from the distribution before and after each update can be measured using relative entropy (def. 4). What Levy (2008) could show is that, if one only takes sentential contexts into account, the relative entropy after each input word is equal to this word’s surprisal (p.1131). Since surprisal is only based on context (other words, in the intra-sentential case), this has the astonishing consequence that the effects of two different latent structures (i.e. linguistic models) are indistinguishable from each other. In other words: as long as two models induce the same contextual probability distributions, it does not matter which model we choose (at least as far as surprisal is concerned).

This result is rather remarkable, as it shows that even simple models can yield appropriate predictions if they come with a well-defined and consistent estimation procedure. Moreover, Levy (2008) argues that his insights also relieves the interlocutors of having to keep track of probability distributions explicitly. Rather, because they need to update their construal of the sentence in some way, the calculation of surprisal falls out automatically from the update process. This means that not only is surprisal an important predictor in psycholinguistic models, it is also likely to correspond to some real process underlying speakers’ and hearers’ parsing strategies. I will outline one such process that has recently been suggested in Section 3.3.

To conclude this section, let us turn to the super-sentential level. Genzel and Charniak (2002) and Genzel and Charniak (2003) attempted to estimate how the information density of sentences in newspaper articles and novels developed as these went on. For this purpose they tracked the average log n-gram probability per sentence using trigrams\(^{19}\). The figures they present seem to use surprisal, however, since all values are given in bits (i.e. they are positive). The hypothesis of Genzel and Charniak (2002, 2003) was that average surprisal should increase throughout the text. The reason is that all surprisal values only take intra-sentential contexts into account. However, the reader will accumulate contextual knowledge from foregoing sentences, as well. Hence, if the information density should stay constant throughout the text, later sentences within a paragraph should be less predictable when looked at in isolation. Nevertheless, they should be equally predictable as all other sentences, given prior

\(^{19}\)The authors themselves claim to have estimated the sentences’ entropy. This is however not possible since entropy is defined only for probability distributions (def. 3). In fact, if one views the language production as a stochastic process, a sentence is the outcome of an experiment with that stochastic process.
Figure 3.6: Average surprisal per sentence in a paragraph. The estimates are taken from newspaper texts. Although the y-axis is labelled *entropy estimate*, surprisal was the value actually used. The solid and dashed lines represent parser and trigram estimates, respectively. Graphic taken from Genzel and Charniak (2002).

context. Unfortunately, there are no known methods to reliably estimate such context (Bell et al., 2009). Therefore, Genzel and Charniak (2002, 2003) opted for the rather implicit approach of looking at isolated surprisal values. Their findings seem in line with their hypothesis (see Figure 3.6). Notice that these findings do support but not confirm the view that texts are arranged so as to keep information density stable through a paragraph. This is because of the indirect estimation method used by the authors. On the other hand, taking into account what we know from the studies reported above, there is little reason to believe that this should not be the case. It would indeed be strange if language followed different functional principles on different levels of granularity.

Apart from the general trend towards higher information density in later sentences, Genzel and Charniak (2002) also found again a repetition effect for nouns. To this end, the n-gram estimates were interpolated with an estimate of how likely a particular noun was to occur based on how often it had occurred in previous sentences. This interpolation decreased the surprisal for nouns. In Genzel and Charniak (2003) it has furthermore been shown that surprisal
at paragraph starting sentences is usually lower than for sentences within the paragraph. Again, this is in line with a theory of constant information density. Sentences at paragraph beginnings have relatively little contextual information, hence their surprisal when viewed in isolation should be lower. On top of that, Genzel and Charniak (2003) also looked at literary writing. They found generally higher surprisal values than for newspaper text. Both the last mentioned findings are interesting not only from a linguistic standpoint but also from a computational linguistics one. After all, surprisal might be used to automatically detect paragraph boundaries and assess literary quality or complexity. This points to many possible future applications of information-theoretic techniques to linguistic work in the broadest sense.

At this point we have reviewed the application of information-theoretical notions to very diverse linguistic phenomena at different levels of granularity. Hopefully, this suffices to convince the reader that many aspects of language are indeed better described and accounted for using statistical techniques than rule-based systems. Nevertheless, we have only considered language processing so far. In the next section I will present a theory of associative learning that can to some extent explain how these processing effects might arise from a general learning mechanism. As the reader will realize, this is in contrast to theories that (i) view learning and processing as independent processes or (ii) assume different processing modules for different phenomena.

### 3.3 Learning

Language learning has for a long time been one of the most puzzling challenges to linguists. It has frequently been assumed that for humans to be able to learn languages, some linguistic knowledge has to be innate. This assumption is grounded on the poverty-of-stimulus argument, which claims that children do not receive enough input to correctly learn a language (for the latest incarnation of the poverty-of-stimulus argument see Berwick et al., 2011). In particular, there are certain rare constructions that they are unlikely to encounter often enough to master them. Moreover, language users have the capacity to produce and understand previously unheard sentences. Unfortunately, what part of linguistic knowledge is supposed to be innate has never been specified by proponents of the theory. Lexical items are unlikely to be innate as this would mean that humans were born with all lexicons of all languages available. Apart from the fact that this by itself would seem absurd, it would also not explain interspeaker vocabulary differences. Another variation on the theme of innatenes are concepts. Intuitively, these are abstract representations of the things around us. Again, it has never been specified what they actually are (see also Ramscar, 2010). Moreover, even if concepts were innate, this would still not explain how learners manage to make the link between concepts and lexical items (see also Ramscar et al., 2011). The most recent innateness claim is that the ability to syntactically combine constituents on different levels of embedding be innate (Hauser et al., 2002). Specifically, this ability is dubbed “recursion”. Notice,
however, that the author never define what they take “recursion” to be. Moreover, levels of embedding are particular to a specific syntactic framework and hence no actual feature of human language. Rather they are assumption introduced by a specific model that operates with specific latent structures. It is therefore hard to see how embedding levels could justify advocating “recursion” as an innate building block of linguistic behaviour, even if the term was defined.

The poverty-of-stimulus argument also neglects that humans and other animals do not only learn from positive evidence but also from negative evidence (Rescorla and Wagner, 1972). This negative evidence may come in two forms. First of all, the learner may find that an event that he predicted based on the cues available to him does not occur. Hence, his prediction is disconfirmed (see e.g. Ramscar and Baayen, 2013; Ramscar et al., 2010, 2011). Second, the learner may produce an utterance to achieve a certain goal (e.g. to be fed). If the utterance does not elicit the desired response from his environment, the learner will have to adjust so as to achieve his goal in later attempts\textsuperscript{20}. In the following we will explore a learning model that is mainly built on the first form of error-driven learning but could in principle also be applied to second one.

\textbf{Naïve Discriminative Learning}

Several decades ago Rescorla and Wagner (1972) proposed a theory of conditioned responses that was built on experiments they conducted using rats and rabbits. By that time it was well known that conditioned responses can be elicited by training the subjects on a conditioned stimulus synchronously with an unconditioned stimulus. An unconditioned stimulus is one that directly evokes the response, for example a mild electric impulse that forces muscle contraction and hence movement of specific body parts. The unconditioned stimulus is one that is not causally related to the unconditioned stimulus and the corresponding response and can be brought into relation with them through observation of co-occurrences. For the remainder of this thesis I will refer to the conditioned stimulus as \textit{cue} and to the unconditioned stimulus as \textit{target}. Both will usually be perceptual stimuli that don’t necessarily need to come from the same domain. The unconditioned stimulus elicits some form of behaviour. Hence, we are looking at a theory of behavioural learning here. This is why it fits nicely with our view on language as human behaviour.

The major insight of Rescorla and Wagner (1972) was that the activation of the response depending on a target can be predicted from the sum of the activations that the cues present would assign to the target. Assuming that any target and any response can only be activated up to a certain upper bound, this has important implications for learning. Given that a target is activated by several cues and the target is indeed present, all the stimuli will be reinforced with respect to that target. If the target is not present, on the other hand, all the cues predicting it will be inhibited with respect to that target (i.e. the association between the target and the cues will be partially unlearned). Cru-

\footnote{I thank Yi Xu for pointing out this possibility to me.}
cially, the level to which the association between cues and targets is weakened or strengthened depends heavily on the association strength that each cue used to have with the target before. For simplicity’s sake let us assume two cues A and B. While the nature of the cues and targets does not actually matter, we may conceptualize A as a dog and B as a horse. We will then take the word dog as the target. We will first take a look at positive reinforcement.

If A alone is observed in combination with the target several times it will gain associative strength with respect to the target\(^1\). This means that the dog will be a predictor for dog. Now if a horse enters the scene, it will receive relatively little re-inforcement with respect to dog. The reason is that the associative values of A and B add up. Hence, even though B, the horse, may have little to no association with dog, the fact that A already has a strong association leads to the compound cue as a whole receiving little re-inforcement. Intuitively, the reason is that the associative strength of the compound is already high enough to (almost) accurately predict the target by virtue of the the compound including A. In other words, because a dog alone is a salient cue for the word dog, the presence of the horse does not figure in much when it comes to predicting that word. On the long run, if A gets re-inforced repeatedly with respect to the target, it will completely annihilate any association between B and the target. This means that the cues are in fact competing for associative strength with respect to the target. This form of cue competition is what allows the learner to make stable predictions. Once a certain cue emerges with a high advantage over its competitors, this cue will be heavily relied on for prediction of the target(s) it is strongly associated with. On a high conceptual level this just means that the learner learns through experience.

The second important part of the Rescorla-Wagner-model is negative re-inforcement or inhibition. Again, assume that A has a strong association with the target, i.e. that a physical dog is strongly associated with the word dog. Then if the AB compound mostly occurs in the absence of the target, i.e. if the word dog is hardly ever mentioned when both a dog and a horse are present, cue B will acquire negative association with the target. The reason is that the compound is overpredicting the occurrence of dog. This is mostly due to cue A. However, since the cues occur as a compound both of them will receive negative re-inforcement. Assuming that A alone gets still re-inforced on occasion and that B never (or very seldom) experiences such a treatment, the only form of association that B will acquire with respect to the target is negative association. Hence, B will be learned as an inhibitor for the target. As a consequence, B alone will henceforth suppress the target dog. Crucially, if we now introduce another cue C and have it occur in combination with B and the target, C will very quickly become very strongly associated with the target. The reason is that the BC compound should in fact suppress the target due to the presence of B. However, since the target is present, both cues will be re-inforced. This re-inforcement will be positive and strong since the actual prediction was far

\(^1\)In fact, Rescorla and Wagner (1972) could show that the acquisition of association not only depends on target frequency but also on the strength of the target. Bigger electrical shocks had stronger effects on learning.
off. Assuming that C initially does not carry any associative value, it benefits a lot from this re-inforcement.

At this point, it is important to notice that both forms of re-inforcement are forms of error-driven learnings. In the positive re-inforcement case the cue(s) underpredicts the strength of the target whereas in the case of inhibition, the cue(s) overpredict the strength of the target. Only when the cue(s) perfectly predict the target will no learning take place. This may seem natural to anyone active in the field of formal learning and machine learning, where the goal generally is to minimize some objective or cost function, which usually is an error measure. It is at odds with any learning theory that claims that learning takes place only on the basis of positive evidence. In fact, negative evidence is just as crucial. An important implication is that learners should strive to make mistakes instead of trying to avoid them. Amongst others, this is supported by finding of Karpicke and Roediger (2008) who could show that in second-language vocabulary acquisition the learning effect was enhanced not through repeated study but through repeated testing. Most likely, it was the error feedback during testing that helped the students to sort out their mistakes. Though Karpicke and Roediger (2008) do make no connection to the Rescorla-Wagner-Model, their findings can be readily explained by it.

Another important and somewhat unexpected consequence of this theory of association is that cue-target associations may be enhanced even in the absence of training examples. This has for example been shown by Ramscar et al. (2011) who provided evidence that when children are trained on even numbers they will at the same time implicitly acquire knowledge about the odd numbers. Crucially, the children were first presented with a picture of a set of known items and then with the number words. This gave them the chance to recognize the cues in the known items and develop predictions about upcoming events, including words. Since texture, colour etc. failed to predict the number words, they would at some point learn the quantity as predictor for number. In addition, they would also learn to dissociate different quantities. Hence, having learned (to an extent) that quantities of two are usually followed by *two* and quantities of four are usually followed by *four*, the children could figure that quantities of three likely won’t predict any of these words. Hence, the search space of possible events following quantities of three was likely reduced. Moreover, all children in the experiment had had some prior experience with quantities and number words. Hence, erroneously acquired associations could be rectified and correct ones could be positively re-inforced. Both the processes of cue competition and error-driven learning are important here.

Before we come to the formalization of the model, one remark is due. I initially introduced cues as conditioned and targets as unconditioned stimuli. While this sticks to the original theory of Rescorla and Wagner (1972), it is a bit to narrow a definition for our purposes. In the following, cues and targets can be any events. In particular, an event may serve as a cue or a target depending on the situation. As a running example we will stick to horses and dogs as cues and *dog* as a target.

Turning to the formalization of the Rescorla-Wagner-model, we see that
it basically consists of a simple update rule, also known as Widdow-Hoff-rule (Ramscar et al., 2011). The reader familiar with machine learning will realize that this rule is the update rule in gradient descent algorithms (e.g. employed in linear regression) and perceptron learning. We will come back to this analogy soon, when we discuss the implementation of the na{"i}ve discriminate learning model. First, however, let me introduce the Rescorla-Wagner-model itself. The update rule is given in equation 3.9 and the formulas for computing the change in weights (i.e. associative strength) is given in equation 3.10.

\begin{align}
V_i^{\text{new}} &= V_i^{\text{old}} + \Delta V_i \\
\Delta V_i &= \begin{cases} 
0 & \text{if } C_i \notin PC \\
\alpha_i \beta_1 (\lambda_t - \sum_{j : C_j \in PC} V_j) & \text{if } C_i \in PC \text{ and } t \in PT \\
\alpha_i \beta_2 (\lambda_{-t} - \sum_{j : C_j \in PC} V_j) & \text{if } C_i \in PC \text{ and } t \notin PT 
\end{cases}
\end{align}

In the above equations, $V_i$ is the associative strength of cue $C_i$ and $\alpha_i$ is this cue’s learning rate. Furthermore, $\beta_{1,2}$ are the learning rates of reinforcer and non-reinforcement, $\lambda_t$ is the target-specific maximum value of activation that this target can achieve if that target is present, $\lambda_{-t}$ is the maximum value if that target is absent and finally $PC$ and $PT$ are the sets of the situationally present cues and targets, respectively. Notice that the overall activation depends on all the cues present. Therefore, in order to discriminate between cues, the learner has to encounter different cues in different situations in order to be able to discern them. Also, because an informative cue will occur more often together with a target than most other cues, it will marginalize these other cues on the long run with respect to the target in question. $\lambda_{-t}$ is usually set to zero since it makes sense to assume that a target that is not present will not have any associative strength (neither positive nor negative).

Although the mechanisms described above were originally designed to model conditioned behaviour, they can be straightforwardly applied to language learning and processing. In fact, under the Rescorla-Wagner-model language is but a conditioned behaviour. This fits squarely with the view of language as human behaviour advocated here. Convincing examples showing that humans likely learn language through experience of co-occurrences of words and other events has been provided by (e.g. Ramscar et al., 2010, 2011). They could show that children learn abstract concepts like colours and numbers better if they are first presented with an amount of known objects or a known object with a particular colour and then with the corresponding number or colour word. The reason is that this way children can focus on the known objects. Since the objects are known the children will likely know a range of words associated with these objects, such as bear and cuddle for teddy bears. Observe that the teddy will predict the occurrence of these words. However, upon hearing the word three or orange these predictions turn out to be erroneous. According to the above equations the associative strength between all cues and the predicted targets
will decrease slightly. At the same time, however, the associative strength between all cues and the number or colour word just heard (i.e. three or orange) will be strengthened. If this is repeated over a variety of scenarios involving different known objects that cue different known words, the quantity or hue will eventually be acquired as cue for the number or colour word. In particular, it will be these features that discriminate between the number or colour word and other words such as bear or cuddle. Notice that this process of re-inforcement is speeded up and strengthened the more often a particular cue-target combination is encountered by the learner. Incidentally, this is what Ramscar et al. (2011) used to explain the relatively faster acquisition of words for smaller quantities. The reason simply is that these quantities are mentioned more often.

The learning model underlying the research reported in Ramscar et al. (2010, 2011) as well as Baayen (2010) and Baayen et al. (2013) is called Na"ıve Discriminative Learning (NDL). Its implications for processing are outlined in Ramscar and Baayen (2013). The model basically implements the Rescorla-Wagner-Equations\textsuperscript{22}. The model is discriminative because it attempts to learn the cues that best predict a given target. These cues will be discriminative for that target. The model is naive in two respects. Firstly, it does not make any assumptions with respect to prior knowledge. Everything is learned from scratch. In its current implementation, the model is even parameter-free since the $\alpha$, $\beta$ and $\lambda$ parameters from equation 3.10 are set as constants which do not vary for different cues and targets. Secondly, the model is naive because it is not capable of drawing any connections between cues that predict the same target (i.e. it does learn semantic similarities). For example, assume that a child grows up bilingually with English and German. Assume furthermore that it only hears one language at a time. Eventually, the child will come to learn both horse and Pferd as cues predicting horses (and the other way around). Naturally, the child will also understand that the meanings of these two words are to some extent similar. The NDL model, however, will not learn this kind of similarity\textsuperscript{23}. In general, the NDL model will struggle to learn that two words are synonymous. It is hence a very basic model that likely suffers from restrictions that do not apply to human learners. Put differently, all associations that can be learned by the NDL model can most probably also be acquired by human learners.

Another assumption underlying the model is that learners distribute their attention equally amongst cues and targets. This is very unlikely to be true since organisms need to filter incoming information and concentrate on whatever they consider most important. Notice, however, that this problem is not specific to NDL but a problem for all current learning models as we do not yet have a

\textsuperscript{22}A recent implementation that works on character sequences can be downloaded from http://cran.rproject.org/src/contrib/archive/ndl/ndl_0.1.1.tar.gz. It was originally designed to predict response latencies in lexical decision tasks but can potentially be used for other applications, as well.

\textsuperscript{23}Since the NDL model is implemented as a perceptron learner, this means that the units in the perceptron are not laterally connected and can therefore not influence each other. Neural networks that do have such lateral connections are known as recurrent neural networks (Elman, 1990). They are very powerful since they can effectively “remember” previous input patterns. Hence they can also establish similarities between these.
way of quantifying attention. Moreover, while this simplifying assumption may lead to incorrect predictions in individual situations, the asymptotic learning outcome should not be affected by it.

**NDL and Information Theory**

It is now time to make the connection between NDL and information theory. It is probably apparent to the reader how NDL can be united with production. The sender chooses a message and an encoding so that he is most likely to be understood by his interlocutors (according to his prior experience). In terms of NDL this means that he chooses a behaviour that is easy to discriminate from other behaviours with respect to the intended communicative goal. But how about comprehension? First of all, notice that under the NDL view there is no inherent difference between comprehension and production. Both are governed by prior experience. This is in contrast with other information-theoretic theories of language such as Jaeger (2010) who explicitly establishes his account as a theory of language production. Notice that Shannon (1948) assumed that the only reason why a physical system would need to process information in the first place is that the system is inherently uncertain about its environment. Specifically, it is uncertain about the message sent by the source. Every incoming bit of information helps to restore the message and hence reduces that uncertainty. Under the view presented here, signals are behaviours and messages are states of the environment. Notice that it is crucial for any organism to be able to predict its environment to a certain extent as it would otherwise not be able to survive. In the case of language, uncertainty is reduced mostly with respect to the (communicative) intentions of the speaker. In the words of Ramscar and Baayen (2013) this presents “a more straighforward view of comprehension as the reduction of listeners’ uncertainty about speakers’ intentions as messages unfold...” Crucially, this includes the left-to-right processing that we encountered in much experimental work.

The connection between NDL and information theory is further supported by findings of Baayen (2010) and Baayen et al. (2013) who could show that many of the contextual predictability effects usually ascribed to conditional probability are in fact highly correlated with NDL weights (associative strengths). However, while an account base on mere frequencies needs to postulate some kind of storage facility for counts and/or probabilities, NDL manages to account for frequency effects straightforwardly through its general learning mechanism. Due to its simplicity and domain-generality it appears that NDL offers a good alternative to theories that only look at frequencies. In particular, it is able to explain why certain effects related to surprisal should arise in the first place. Moreover, prediction plays a crucial role in NDL. In fact, viewed from an NDL perspective processing difficulties do not arise so much because a word is surprising but because other words have been predicted to occur in its place. This is in line with Levy (2008) who also assumes prediction as a crucial factor in language processing. It is at odds with theories like Jaeger (2010) which solely focus on surprisal. Recall that by the proof in Levy (2008) the information-
3.4. ADDITIONAL REMARKS

Theoretical notion of surprisal (def. 2) is compatible with prediction. The issue is hence more conceptual than formal.

Its focus on prediction speaks in favour of NDL, however. The reason is two-fold. First of all, in order to be proactive any organism needs to be able to predict what is going to happen around it. Hence, prediction is crucial in all aspects of life. Secondly, as I mentioned above the learner may also gather negative evidence by exerting a behaviour with a particular goal in mind and this goal may not be achieved. In most simulations with NDL the targets have been taken to be events generated independently of the learner’s behaviour. However, there is no reason why targets could not be events that the learner predicts to happen in response to his own behaviour. Crucially, without prediction such goal-driven behaviour would not be possible in the first place.

In sum, we have a learning-based framework that allows us to cover important aspects of human language use and nicely integrates with findings that suggest an information-theoretic conception of language. Moreover, NDL is domain-general and not specific to language acquisition. It therefore fits with the view that language is a form of human behaviour.

3.4 Additional remarks

At the end of our discussion of the relation between language and the notions from information theory, I would like to turn back and briefly look at more conventional linguistic notions. More precisely, I want to point out what consequences the view outlined above may potentially have for the field of linguistics.

3.4.1 A Note on Grammar

Grammar is perhaps the most pervasive notion in the field of linguistics. Indeed, formal grammars have led to significant advances in the field of computational linguistics (especially in parsing) and also in compiler theory. What often slips the attention of linguists working with grammars is that they consist of fixed vocabularies (both terminal and non-terminal) and a fixed set of rules. Natural languages, on the other hand, are in constant flux. The conventions of their use as well as the vocabularies, sounds and scripts they employ undergo changes all the time. From this formal viewpoint it should be apparent that formal grammars are a particularly bad tool to model the psychological and behavioural processes underlying natural languages. Furthermore, most grammars are designed to impose constraints on possible constructions. They carry no implications as to when and how these constructions should be used. This means that even if one knew the hypothetical grammar of a language, one would still not be able to communicate adequately with the users of that language. Under the view introduced in the previous section, on the other hand, exerting an appropriate behaviour under a given set of circumstances is all that counts. This may e.g. mean to flee when a predator is approaching, to secure a food source once it is discovered and in the case of language it means to behave in
way that the others can make sense of so as to arrive at a successful instance of communication. Importantly, in line with information theory, we regard communication as successful as long as there is a mutual understanding of what is being said. Success is not influenced by whether or not the interlocutors arrive at their communicative goals. Also no cooperativity is required. Instances of communication that attempt to deceive have the same conditions of success as instances in which the speaker tries to be helpful.

In spite of this, the notion of grammar seems to have developed into the holy grail of linguistics. Even authors working experimentally adhere to it for no apparent reason. An example is Jaeger (2010) who accepts and supports the influence of prediction and surprisal on natural language use but at the same time states that only “Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal [...]” (p.25, emphasis mine). This is especially peculiar since in terms of cognitive and processing capacity nobody even knows what “grammar” is supposed to be. It seems that the term is but an empty hull which linguists fill with their preconceptions of how language is supposed to function.

The theory reported in this thesis explicitly goes against this (common) practice. Instead of theorizing on the basis of intuitions and individual judgements, it advocates experimental work and the grounding of theories in experimental findings. Moreover, linguistic theories should strive to uncover the functional principles of language use. The conception of language use as a form of human behaviour provided above is certainly not flawless. To the best of my knowledge, however, it is the most general one and also most consistent with what we know about behavioural patterns of other species. Crucially, it manages to derive all of its modeling power without any appeal to grammar. Recall that for a processing theory based on surprisal, the choice of latent (i.e. grammatical) structures is irrelevant as long as empirically adequate surprisal values can be assigned to words (Levy, 2008). In the absence of a precise definition of natural language grammar and without concrete evidence about the effect on language use and processing of a concept thus defined, it seems advisable to dismiss of the notion of grammar entirely in the context of natural language.

3.4.2 A Note on Semantics

The reader familiar with logic-based semantic theories of focus such as Rooth (1985, 1992) may be left wondering which place is allocated to the meaning component of focus in the picture painted here. The answer is simple: it all depends on what aspects of meaning one is looking at. As outlined in section 3.1, no specific interpretation mechanism is assumed here. Rather, humans likely extract the meanings crucial to them from a complex signal. These meanings may range from factual over social to body-related meanings (a hoarse voice for example may indicate that the speaker has a sore throat). No matter which form of meaning one is interested in, under the theory described here all these meanings come about through discriminative association of meaning manifestations (e.g. the physical or pictorial presence of a horse) and corresponding
signals (e.g. the words horse, Pferd or paard but also competitors like dog). Hence, semantics in the sense of association between events is ubiquitous in this theory.

Some important consequences of semantic ubiquity are that the traditional boundaries between subfields of linguistics, such as syntax, semantics and pragmatics, vanish. In fact, the distinction between them is artificial and has only been imposed to delineate different complementary academic areas of work from one another (Abels and Neeleman, 2010). In contrast to what Abels and Neeleman (2010) are claiming, it seems that the time (and the technical apparatus) has come to dispose of these artificial distinctions and aim for more encompassing theories of language.

Notice that in any theory according to which linguistic signals (and also non-linguistic signals, for that matter) are learned discriminatively, the division of different layers of meaning within the signal does not make sense. This is because a) signals by themselves have no meaning and b) multiple meanings may be encoded in the same signal. From the perspective of the receiver it only matters whether the signal can be decoded into a message. It is the target (or the recipient in human communication) that has to make sense of the decoded message and discern different forms of meaning. In sum, semantics is a very important aspect of the theory advocated here. If we view semantics as the connection of two or more events, semantics occurs everywhere in language. Hence the notion of semantics employed here is much richer, but also less specific, than any other notion of semantics commonly found in the literature. Moreover, the traditional boundaries between syntax, semantics and other allegedly separate parts of language are blended smoothly together.
Chapter 4

Focus and Information

The present chapter proposes a new account of the linguistic function of focus. In a first step, focus is analyzed in information-theoretic terms. I argue that focus has, amongst other things, the function to lengthen words with high surprisal values. In a second step, I point to the social functionality of focus. As a conventionalized linguistic tool focus likely helps speakers to coordinate their discourse interaction. A breach of these expectations can hence be a breach of social conventions at the same time. This claim will be backed up by experimental findings elicited in a large-scale online study.

4.1 Focus Duration and Information

Now that I have given a thorough overview over state-of-the-art research results regarding focus and the connection between language and information-theoretic notions, the time has come to integrate the two. Recall that in Section 2.1.1 we established pitch range expansion, higher durations and post-focus compression as three of the most important and most consistent cues to focus coding in English. Notice that languages which employ phonetic focus marking do not necessarily make use of PFC. They do use the other two cues, however.

In the previous chapter we also saw that surprisal is one of the most significant predictors for word durations in speech (e.g. Bell et al., 2009; Gahl, 2008). The more surprising a word in its context, the longer it tends to be pronounced. According to the working definition of focus given in def. 1, focus in statements is intimately tied to questions. An example of a dialogue that would likely involve focus is given below.

(1) A: What did you have for lunch yesterday?
   B: I had carrots for lunch yesterday.
   B’: Carrots.

Either version of B’s answer will be fine, the first one perhaps being fairly redundant. As we know from empirical studies, the word carrots will be focused
in any case (Cooper et al., 1985; Eady and Cooper, 1986; Xu, 2005). Since A asked the question in the first place, he is likely to not know the answer. This in turn means that the element that actually answers his question, i.e. *carrots*, will be rather surprising in this context. Put differently, the information density on this word will be high.

A proposal that I have recently made (Schulz, 2013) directly builds on the above observations. Since focus lengthens the word it marks, it helps reducing the information density on that word. In terms of processing, it gives the interlocutors more time to integrate the word with their current parse. Hence, focus supports the discourse flow by making surprising words easier to integrate. This proposal has potentially large explanatory power as it directly connects the phenomenon of focus to the well-established findings about more general linguistic behaviour and its processing. In particular, it makes the connection between the physical properties of the signal (Section 2.2.2) and information theoretic measures (3.2). It hence lessens the need to linguistic theories that are particularly tailored to the phenomenon of focus. Moreover, this proposal opens up a new perspective on information structure, which is arguably more integrative than the theory of information structure currently employed (see e.g. Krifka, 2008). It takes the term for real as it views focus as structuring the information flow as quantifiable for example through surprisal. However, this view of information structure can be applied to all forms of linguistic behaviour, leading to a ubiquity of information structure. This is in line with research based on ideas of the NDL model (Ramscar et al., 2010, 2011) where it has been shown that information structuring in terms of feature-label ordering is crucial for learning.

Before I further explicate some of the benefits of my proposal, allow me to first address potential criticism that it might receive.

**Potential Criticism**

The first point of contention that a reader familiar with the phonetics of focus marking may bring up is that the above proposal narrows in too much on the duration part of focus marking. As we have seen in Section 2.1.1, there is a variety of other focus markers, as well. This is true and my proposal is not intended to sideline these other focus markers. Notice therefore that I do not claim that lengthening for processing enhancement and information density spreading be the only function of focus. On the contrary, I hold that it is but one of a multitude of functions. At the same time, it seems to be the one that is best to explain in general processing and information-theoretic terms. While in isolation it admittedly constitutes a poor account of focus, it will likely be very powerful if other factors are taken into account. These factors include memory enhancement (Gernsbacher and Jescheniak, 1995; Birch and Garnsey, 1995) and arguably also attention attraction.

A second point that may raise criticism is again related to focus marking. After all, focus has been defined in functional terms in this thesis (def. 1). Nevertheless, not all languages use phonetic focus marking, but instead rely
exclusively on morphosyntactic markers (Hartmann and Zimmermann, 2009). Still these languages exhibit focus according to def 1. This appears to be at odds with durational variation due to focus. Notice, however, that in such languages there usually are specific focus morphemes that are attached to the word under focus. Hence, the word gains an additional morpheme which mostly corresponds to an additional syllable. This in turn leads to a longer pronunciation of the focused word. Hence, there is no obvious reason why languages that exclusively use morphosyntactic focus markers should not employ the same general functionality of focus proposed here.

Thirdly, some readers may be worried about the general correctness of the proposed theory. For example, one may claim that focus does not need to alleviate uncertainty about an answer in all situations. Consider (2) as an example.

\begin{align*}
(2) \quad & A: \text{Did you drink whisky or wine?} \\
& B: \text{I drank whisky.}
\end{align*}

In this case, A explicitly limits the choice of possibilities. Hence, his uncertainty about the answer may not be too high. Arguably, this is true. Recall, however, that we have tied the uncertain nature of language use to experience-based learning through the NDL model. In the case of focus as I have defined it here, this means that the event of a question will be strongly associated with the event of focus. This association will be re-inforced every time a question-answer pair occurs. Thus, though the information spreading function of focus may not be crucial here, focus will still occur simply because it has been learned as a highly informative cue for answers. Furthermore, it bears reiterating that information spreading is not the only function of focus and other functions, such as memory enhancement and attention attraction, may still be at play here. More generally, strong cues will elicit the associated target behavior under the NDL model even when this behavior is not strictly required in a given situation\(^1\). This happens because the target behavior is necessary most of the time (as the learner has come to understand through experience). It would indeed be very inefficient and even dangerous to judge the cue within its current context, if this cue has been learned as a strong and reliable predictor. For example, if a knife falls off a table most people would likely try to get out of its way without caring about how sharp the knife actually is (it might not be harmful, after all). This behavior is the safest since any deliberation about how sharp the knife actually is would postpone the appropriate reaction so an eventual cut could not be avoided anymore. Crucially, we have learned that most knives happen to sharp. The reader is invited to compare this to a similar situation in which a napkin is falling from the table. In language, as well, certain communicative behaviors are exerted without further re-assessment because they are appropriate by experience. Incidentally, this is in line with research in phonetics where focus has been proposed to be a highly stylized device, as well

\footnote{In fact, it is one of the foundations of any theory of conditional learning that conditional behaviors are exerted in reaction to a stimulus without deliberation.}
Finally, even the reader convinced by all arguments so far may still argue that the theory of focus proposed here is too narrow. First of all, the definition of focus (def. 1) may be felt to be to limited and also the closely scrutinized function of duration within focus may be seen as arbitrary. With respect to the first point I would like to point out that our definition has been intended to be a working definition from the beginning. While it is certainly worth extending, I commit to the claim that any reasonable definition of focus will have to subsume the one given here. In that sense, it is a crucial building block. Nevertheless, it should be noted that currently there simply no better definition. I have pointed out weaknesses of theories driven by semantic and phonetic considerations in the beginning. Virtually all alternative definitions in the field of pragmatics include some reference to Questions Under Discussion (Roberts, 2012). They extend our definition to cases where no overt question is present. Crucially, as of yet no algorithm to discover Questions Under Discussion in a principled way has been formulated. Furthermore, there is no empirical evidence that they have any cognitive reality. They are therefore solely a theoretical construct that can currently not be put to any use. Consequently, any focus definition built on them will suffer from the same problem.

The reason for concentrating on effects of focus duration was two-fold. First of all, it was the most accessible aspect of focus marking that could most easily be integrated with work on human language processing in general. In addition, there is good evidence that duration is the most salient part of focus marking. Arnold et al. (2012) report that they found duration to be the acoustic cue that was most strongly correlated with perceived prominence. Prominent words had been elicited through question answering, hence fitting our definition of focus. Furthermore, Watson et al. (2008) could verify that pitch but even more significantly duration are negatively correlated with predictability whereas intensity (i.e. loudness as observed on changes in air pressure) correlates positively with importance. In order to arrive at these results, the authors had their participants play tic tac toe. Earlier moves in the game were construed as less predictable and less important since many cells are still unoccupied and no win can yet be achieved. Later moves were classified as more predictable and more important as winning is possible with one stone whereas the move is rather predictable from the situation on the board. Assuming this classification is right, the authors could deduce their results from the fact that earlier moves were reported with longer pronunciations and later moves with higher intensities. The authors do admit that both measures may be confounded with emotions coming up in a win or loss situation. This is possible, since especially pitch has been shown to be affected by emotions (Chuenwattanapranithi et al., 2008). Summing up putting focus duration under closer scrutiny has by no means been an arbitrary choice but can be well motivated with conceptual and empirical considerations.

\(^2\)Readers unfamiliar with the game should consult the following online description: http://en.wikipedia.org/wiki/Tic_tac_toe
Benefits of the Theory

I will now summarize some benefits of the theory and also connect it again to some empirical evidence. Notice first that most counters to criticism from the foregoing section do in fact constitute support for the proposal. This includes extendability to languages employing only morpho-syntactic focus markers, focus use based on experience-based learning and the general significance of duration for focus. However, the theory also fits with the findings by Jaeger (2010) and Levy (2008) regarding syntactic variation and in particular syntactic reduction. In fact, this kind of variation also exists in answers. Reconsider example (1). B could just as well have said I had carrots. Between the three alternatives there are probably more forms of gradual syntactic reduction. In line with Jaeger (2010), this reduction may be in part explained by information density reduction. The more predictable the answer as a whole, the shorter it will be. Since most parts of an answer other than the word corresponding to the wh-element in the question are highly redundant, it should come as no surprise that speakers often just use a single word answer like carrots. There are, of course, also other factors at play here. Besides the ones reported in Jaeger (2010) these are mainly social factors as short answers may be perceived as impolite in some contexts.

Another benefit is that through the intimate connection to a learning model the proposed theory may to some extent explain the stylized nature of focus marking. It is simply learned through co-occurrence in question-answer contexts. This can obviously not be the whole story since focus occurs in other contexts, as well. This does not speak against our learning theory, however. After all, we may just have overlooked what the actual cues that elicit focus. Conceivably, there may be a multitude of such cues occuring in different contexts. In any case, focus acquisition can be embedded into general language learning under the view proposed here.

Another potentially far-reaching achievement of the theory presented here is that it offers a new view at information structure research. First of all, our definition of focus is functionally driven. This removes the arbitrary component that definitions solely based on focus marking suffer from. Under the present view, every information-structural notion should be defined in terms of its function(s). This gives much more substance to information structure research itself, since it specifies why information structuring is beneficial to language users. Moreover, since we have defined very concrete information measures, we have also specified what kind of information we are talking about. Because our notions of information are purely quantitative, no qualitative hierarchy of different information statuses needs to be established. This removes the problem of constantly having to probe, negotiate and modify such hierarchies. Compare, for example the development of hierarchies from Prince (1981) to Baumann and Riester (in press) in order to appreciate how severe a problem this is.

Finally, since the theory embeds focus into the wider framework of expectation-based linguistic behaviour, it lifts some restriction off focus research. Focus does not need to be seen as a solely semantic or syntactic phenomenon anymore but can be conceived of more generally as a stylized device that humans use in so-
cial interactions. This means that the question of when and why focus occurs can be extended from purely semantic or syntactic concerns to social ones (and possibly others). Whether and to what extent focus helps to establish discourse coherence as well as social coherence through indication of friendliness and sincerity has been investigated in an experiment which is reported in the following section.

Overall, looking at focus from an information-theoretic perspective seems very promising. Making the connection between focus effects on duration and expectation-based language use delivers straightforward explanations for why focus occurs in question-answer contexts and justifies our working definition.

Contextualization

To the best of my knowledge, I have been the first one to explicitly point out the connection between focus-related duration variation and surprisal (Schulz, 2013). However, I have certainly not been the first one to realize that focus is strongly context dependent. In a seminal paper Bolinger (1972) first brought up convincing examples for why accent in general and focus in particular cannot be predicted by a strict rule-based system\(^3\). In particular, Bolinger (1972) pointed out that while there may be a general metric preference for accent placement, it usually happens that words that are rather unpredictable from their foregoing context become accent bearing, thereby overriding that preference. Furthermore, he also observed that emotional highlighting may have an effect similar to focus and also lead to accent placement on a word (see also Chuenwattanapranithi et al., 2008). Bolinger ends his paper stating that while certain syntactic patterns may indeed exert a a preference for a particular intonational contour “a description along these lines can only be in statistical terms” (p.644). Unfortunately, Bolinger himself did not have the statistical means to prove this claim. Modern research methods seem to indicate that he was right, nevertheless.

A second related idea is the smooth signal redundancy hypothesis by Aylett and Turk (2004). They too propose that contextually predictable words should be pronounced shorter. In line with Jaeger (2010) they assume that this serves to keep a stable level of information density or, as they put it, a smooth redundancy profile throughout the utterance. However, they attribute this effect mainly to prosody. Through prosody manipulation, durational variation should be achieved. For focus this would imply that duration is a mere side effect of the pitch movement that focus induces. Hence, we should have looked at pitch instead of duration by itself. Undeniably, there is certainly an interaction between pitch movements and duration (Xu and Sun, 2002). However, the concentration on pitch alone is likely to narrow. First of all, Aylett and Turk (2004) explicitly assume that “Aspects such as structural relations between syntactic constituents […] do not necessarily affect language redundancy …” (p.54). This is of course at odds with several findings about the role syntax plays in

---

\(^3\)Bolinger (1972) in fact never used the term focus but his examples strongly indicate that this is what he was thinking of when writing the paper.
word surprisal. In fact, by its very definition which includes context, surprisal depends on syntax. Secondly and most importantly, it is far from settled that pitch controls duration. Recent research by Xu and Wang (2009) suggests exactly the opposite. They examined pitch contours within syllable groups in a controlled experimental setting with varying syllable positions and speaking modes. They proposed that duration is used in order to make it possible to approximate underlying pitch targets. Indeed it would be hard to explain why pitch patterns, which have been shown to be communicatively functional (Xu, 2005) would be introduced arbitrarily just to extend durations. I therefore side with Xu and Wang (2009) in my own proposal.

Finally, it may appear that there is a connection between the proposal made here and the one by Beaver and Velleman (2011). In my opinion, however, this is not the case. Beaver and Velleman (2011) attempt to establish a relationship between focus, predictability and importance. However, they basically equate focus with importance. This is somewhat unexpected since they cite the study of Watson et al. (2008), who found that the basic phonetic characteristics of focus are mostly elicited in unpredictable rather than important contexts. Moreover, predictability and focus are inherently unrelated notions in the work of Beaver and Velleman (2011) that happen to occur together quite often. The authors do also not define what they actually mean by predictability. It does become clear, however, that they think of a dichotomy between predictable and unpredictable contexts and hence a qualitative one. This is in stark contrast to what has been proposed in the present work. In order to check whether an utterance is in accord with their notion of predictability, Beaver and Velleman (2011) use the idea of giveness from Schwarzschild (1999). Hence, predictability is equated with being deducable (in a logical sense) from prior context. The authors correctly remark that this is only a checking criterion and does not offer a theory of when focus (or “accent” in their terms) should occur. Instead of trying to develop such a theory they retreat to advocating “a few rules of thumb” (Beaver and Velleman, 2011, p.1677). In sum it seems that there notions of predictability and importance are not well-defined (or at least not straightforward to implement). Much of their proposal does indeed seem very ad hoc and not well justified. Hence I do not wish to connect my account to theirs but rather to separate the two from each other as much as possible.

4.2 Focus as Social Behaviour

This section reports an experiment that was designed to test the impact of focus on social interaction. Following def. 1, focus is bound to occur in answers. This likely leads speakers to expect a focus in an answer since they have experienced this pattern many times before. The reasoning behind the present experiment is that although an answer may be satisfactory from a factual standpoint, it may still be perceived as impolite or insincere based on the placement of focus. Consider example (3).
In most cases, one would expect the focus to fall on wine. This is also what def. 1 would predict. However, speaker B might also choose to place the focus on I, for example in order to delineate himself from other people. The factual content of the answer stays the same. Nonetheless, there is a violation of expectations taking place here. This violation can also be formulated in terms of alignment. We may conceptualize focus as a device that helps speakers to align their utterances in the sense that it indicates when the question has been answered. Other such devices are e.g. pauses that are crucial in order for speakers to know when their interlocutor has finished and they can start speaking again. Clearly, such alignments are strongly related to social behaviour. In principle, there is nothing that would prevent speakers from giving arbitrary answers or interrupting each other all the time. There are, however, social conventions that prevent this. Our experiment tested three dimensions of social judgements. Our subjects were asked to rate question-answer dialogues according to how friendly and sincere both speakers were and how confident the answerer was in her answer. Friendliness is a basic trait in social interaction, signaling a non-aggressive attitude. We included it in the present study to investigate whether misalignment of and the breach of social conventions that comes with it would be perceived as less friendly. This variable can be seen as the most important one for our purposes. Friendliness is such a basic trait of human social behaviour and its conception does arguably not vary too widely within a cultural community. Therefore any effect on that variable would soundly establish an impact of focus placement on social aspects of human communication. Sincerity was included because subjects might perceive misaligned dialogues as indicating that the speakers are not very serious about their topic and might just be joking with each other. Finally, we also asked our subjects to rate the confidence of the answerer since misalignment of focus might indicate that the answerer had problems formulating his answer and hence ended up misaligning the focus unintentionally.

4.2.1 Method

We used crowdsourcing to recruit our subjects. The service we employed was www.crowdflower.com. This site allows one to design a task and then make it available on different so-called channels. These channels are other crowdsourcing websites that have a base of workers available to them. Hence, crowdflower has the potential to attract a large number of subjects in a small amount of time since it essentially combines the workforces from several crowdsourcing websites. Crowdsourcing is a fairly recent way of recruiting subjects for experiments and it is obviously less controlled than experiments in the laboratory. After all, subjects are not under the experimenter’s supervision. For Mechanical Turk, the most popular crowdsourcing website, there have been some investigations about its reliability and quality. Paolacci et al. (2010) recruited 1000 workers and elicited demographic data from them. They found that the average Me-
Paolacci et al. (2010) also replicated three well-known tasks from the field of decision making and found that the results they got were close to those acquired in their labs and in any case far closer to the lab results than results obtained by posting the link to the experiments in internet fora. The authors identify the fact that unsupervised subjects tend to be less attentive as a potential problem of crowdsourcing experiments. At the same time, however, the absence of an experimenter prevents any confounds that the experimenter might introduce. Also, since the subject pool is not restricted to university students “U.S. workers on Mechanical Turk are arguably closer to the U.S. population as a whole . . .” (Paolacci et al., 2010, p.413), hence giving the data more ecological validity. Another investigation of Mechanical Turk found that the payment level of the crowdsourcing task, which is generally very low, does not seem to affect the results obtained (Buhrmester et al., 2011). Over all, it appears that crowdsourcing is very well suited to cheaply recruit a large amount of subjects in a short span of time. However, Paolacci et al. (2010) point out that the workership on Mechanical Turk is changing constantly, with more and more people from Asia doing the work. For the present experiment we restricted possible subjects to those with IP addresses in the U.S.

Two female native speakers of American English were recorded in a sound-proof booth at the University of Tübingen. They both produced questions and the corresponding answers. The items they had to produce are given in (4).

(4) a. A: What did you buy yesterday?
   B: I bought some wine yesterday.
b. A: What did Peter buy yesterday?
   B: He bought some wine yesterday.
c. A: Who did you meet yesterday?
   B: I met Helen yesterday.
d. A: Who did Peter meet yesterday?
   B: He met Helen yesterday.

The speakers were instructed to read out each question in a monotone, almost depressed voice. This was done in order to establish a baseline without any focus in it. Note that the speakers first read the questions and then the answers. This way they knew the question and could place the focus appropriately when asked to do so. Focus was put on each of the words except the determiner some and the final word yesterday. The final word was only introduced in order to avoid sentence-final focus which is notoriously hard to detect for listeners. Also, we refrained from focusing quantifiers as these do not appear in all items. Finally we varied whether the question would be about the person asked or another person, in that case Peter. The reason is that we expected that our confidence measure might always be rather high when a person is giving information about herself. To check whether this would introduce any confounds we chose to also have a condition in which the topic of the conversation would be another person not taking part in the exchange.
After having pronounced all questions and answers in a monotone voice, our
speakers were instructed to accent each of the focusable words (i.e. all words
except *some* and *yesterday*) in turn. All question recordings were then combined
with the appropriate answers. This was done joining a question from one speaker
with the answer from the other speaker using audacity\(^4\). In between the offset of
the question and the onset of the answer a pause of 400ms was introduced. This
was done in order to exclude any effects based on answer latency. This gave us
a total of 160 items (2 speakers × 4 question-answer-pairs × 5 question pitch
patterns × 4 answer pitch patterns). For each of these we obtained judgement
from 5 subjects. Each subject only heard one item. We asked the subject to
rate the answerers and questioners friendliness, the answerers and questioners
sincerity and the answerers confidence in her answer on a scale from 1 to 8, where
1 corresponded to “not at all friendly/sincere/confident” and 8 corresponded
to “very friendly/sincere/confident”. The scales were chosen to have an even
number of possible judgement values so as to force subjects to commit to one
direction or the other. We also introduced questions asking for the gender of
the first and second speaker. Notice that both speakers were female so it did
not actually matter which speaker spoke first. These two questions were rather
intended to check whether our subjects had in fact listened to the dialogue. They
served as an exclusion criterion at the evaluation stage. In order to avoid any
question ordering effects the questions were randomly permuted for each subject
using javascript. Our subjects were paid 7 cents each for their participation.

### 4.2.2 Analysis and Results

Besides the target variables (i.e. the judgements of the subjects) the data con-
sists of the following pieces of information which we will use as predictors in our
analysis.

**Questioner** A feature indicating which of our two speakers asked the ques-
tion. It follows that the other speaker assumed the role of the answerer. This
predictor has been introduced to control for speaker-specific effects.

**Object** This feature codes the animacy of the object. Animacy effects are
known to influence for example dative alternation in English (Baayen, 2008,
ch.2).

**Topic** A variable coding the topic that the inquiry is about. This is either
the person being asked or another person referred to as *Peter*.

**Focus location in the question (QFoc)** This variable encodes the posi-
tion of the focus in the question. It has five levels, one of which is the no-focus
condition.

\(^4\)http://audacity.sourceforge.net/
Focus location in the answer (AFoc) Similar to QFoc but for the answer. There are only 4 levels, since the answer has one focusable word less than the question.

For the focus location in the question, recall that we defined focus in questions as the point were the pitch rises onto the question level. The phonetic properties of focus in English are explained in more detail in Section 2.1.1.

For the analysis of our data we first of all had to exclude about half of our subjects (374 out of 800) because they had failed to correctly answer at least one of the two control questions. This is rather unexpected given that Paolacci et al. (2010) obtained completion rates of 91.6% for their experiments with only 4.17% of their control questions incorrectly answered. We reckon that the low reliability of our subjects is due to two factors. First of all the population of crowdsourcing workers has likely developed over the past three years. People got better adapted to the system and potential defrauders have likely found ways to exploit the system. More importantly, crowdflower uses many different services, only one of which is Mechanical Turk. Mechanical Turk allows the employer to control the quality of the results obtained. If the results turn out to be unsatisfactory upon inspection, the employer may refuse to pay the worker. Therefore workers likely feel a strong inclination to do their job well. Crowdflower does not allow for such a possibility. Workers are paid as soon as they complete a task. Poor quality of the work will have no repercussions. All the analyses conducted in the following are hence based on the remaining 426 subjects.

We fitted several linear models to test for effects of our predictors on friendliness, sincerity and confidence. Since each of these response variables was rated on an 8-point scale, we can conceptualize them in different ways. First of all, we can think of our response variables as continuous variables. This is de facto untrue since our subjects did not have a continuous choice. However, it may be adequate for modeling purposes. Alternatively, we may conceptualize our response variables as random variables over discrete classes. This would demand a classification treatment. However, we do simply not know whether our subjects perceived the scores as discrete or conceptually overlapping. In other words: they may have thought that their is some commonality between the 7 and the 8 on each scale because both indicate that the speakers’ utterances were perceived as rather friendly/sincere/confident. In the present case we chose to fit linear regression models only. The reason is twofold. First of all, since we do not know how our subjects conceptualized the classifications we offered to them we assumed that at least between subjects there likely was an overlap in how the scale values were perceived. Moreover, using a classification algorithm may turn out to be cumbersome since we are dealing with 8 classes here, one of which is much more frequent than the others. A classifier may easily be led astray by this and focus too much on the dominant class. Also in order to

5The reason why we chose crowdflower in the first place is that it is virtually impossible to obtain a Mechanical Turk account in Europe.
Figure 4.1: Barplots showing the distribution of judgement values with respect to each of our social variables. In the Confidence plot there are no values for the questioner as this variable has only been tested for the answerer.

train a classifier on 8 discrete classes the amount of data at our disposal may simply be insufficient. If we interpret our response variables as continuous, the sparsity issue becomes much less severe. All the models were built using the R statistical computing software (http://cran.r-project.org).

Before we start the actual modeling process, it is worthwhile to look at how our data is distributed. Figure 4.1 shows how many times each score was chosen for the three social variables that we tested. We clearly see that the ratings are skewed towards the highest possible value. Moreover, the friendliness and confidence ratings are bimodal with a first peak around 3-4. Incidentally, these are the highest values subjects could give in case they thought that the speaker
was rather not friendly/sincere/confident. Hence, no matter whether subjects agreed or disagreed with a possible ascription, they tended to choose the values that were most favourable for the speakers they listened, too. Once speculative reason for this might be that people in the U.S. are socially trained to not be upfront. That means that even in a situation where they are not directly confronted with the person they have to judge they still prefer to comfort that person instead of potentially offending them. However, I will not delve deeper into this issue here as our data does not allow us to make sound conclusions about it.

Linear Modeling

To approach the modeling of our data we chose linear regression models. These are widely used for data analysis and very well studied. Other models, such as linear mixed effect models, build on them. In fact, the fixed effect in a mixed effect model are those of the standard linear regression model. It would certainly exceed the purpose of this thesis to give a thorough introduction to regression modeling. The reader may intuitively understand a linear regression model as a model that tries to establish a linear relationship between the predictors and the response variable in \( n + 1 \)-dimensional space where \( n \) is the number of features and one dimension is reserved for a bias term. To achieve this the model tries to estimate the \( \beta \)-coefficients in a linear combination of the form given in equation 4.1. The \( x \) values are the input features with \( x_0 \) being the bias constant 1.

\[
y = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_n x_n
\]

We first fitted models for the two friendliness ascriptions. In the following I will only report the final models. However, the reader should be aware that the general modeling strategy consisted of first introducing the control variables (topic, object, questioner) before introducing our main variables. The significance of a predictor was assessed using Anovas (see Jaeger, 2008). An Anova compares two model with and without a specific predictor. It also penalizes models that use more parameters than others. In the following I will only report models that with significant predictors. The reader should be aware, however, that we followed a modeling strategy according to which we first introduced the control variables (object, topic, speaker) and then the main variables. If a variable turned out to be insignificant it was subsequently removed from the model.

We did not obtain any significant results for sincerity ascriptions to the answerer. Also the answerer's confidence was only marginally influenced by the identity of the speaker, with speaker 2 being perceived as slightly less confident. For our main target, friendliness, we did find significant effects of focus, however. As can be seen from Table 4.1, the answerer was perceived as drastically more unfriendly if, in response to a question with pre-final focus, she put focus on the second word. In our experiment the pre-final word in the question and the second word in the answer were always the verb. In a way, this means that the answerer was mimicking the intonation of the questioner to an extent. Apparently, this kind of behaviour is perceived as very unfriendly. Table 4.1
4.2. FOCUS AS SOCIAL BEHAVIOUR

Table 4.1: Regression results for answerer friendliness ascriptions. The upper table contains the anova results, the lower one is a summary of the fitted regression model. SQE stands for squared error. The Df columns contains the degrees of freedoms (parameters) used by each predictor. Predictors and the corresponding factors that are separated by a colon indicate an interaction. The focus position is indicated by foc followed by the number of the word on which the focus was placed. The reference level for all focus-related effects was the utterance which did not contain any focus.

Table 4.2: Linear model for friendliness ascriptions to the questioner.

Table 4.3: Linear model for sincerity ascription to questioner.
Table 4.4: Summaries of logit models. The upper one corresponds to the linear regression model in Table 4.2, the lower one to the one in Table 4.3.

Table also shows that speaker 2 was perceived as far friendlier than speaker 1 when giving the answer. This is in contrast with the confidence ascription according to which speaker 2 was perceived as less confident. While this may hint at an inverse relationship between friendliness and confidence, further research would be needed to establish such a relationship reliably. According to the model in Table 4.2 it seems that answers are perceived as less friendly when the focus falls on the wh-element or the verb. For the questioner’s sincerity we also find a strongly negative effect of pre-final focus (on the verb).

A remark concerning the quality of our models is due. As was to expect, they suffer from a rather low fit to the data. This is, however, not so much the model’s fault as the data’s. After all, the model tries to make continuous predictions on discrete data. However, we also binarized the response variables, joining values 1-4 and 5-8 together and then fitted logistic regression models. Although these models, which make categorical predictions, had a much higher fit to the data as indicated by the AIC\(^6\), they by and large confirmed our findings from the linear models. Table 4.4 shows the summaries for the logit models corresponding to the linear regression models in Tables 4.2 and 4.3. For the questioner’s friendliness the logit model introduces a new negative effect for the second speaker. For the questioner’s sincerity the logit model dampens the animacy effect. For the friendliness ascriptions to the answerer a logit model yielded no significant effects. When it comes to interpreting the coefficients of the logit model, the directionality stays the same as with ordinary linear regression models. Note, however, that logit models are linear in log-odds space but non-linear in probability space. Hence, the coefficients cannot be interpreted as linear coefficients with respect to the probabilities of an item falling into one

\(^6\)The AIC (Akaike Information Criterion) is a measure of model-likelihood that at the same penalizes for the number of parameters. In contrast to the Anova, it does not require the models it compares to be nested. This makes it a convenient tool for comparing models with different model structures. In particular, it allows us to directly compare the linear with the logistic Regression models. The reader interested in the AIC may consult an excellent introduction in Burnham and Anderson (2004).
class or the other.

We take the results obtained from logit modeling as a confirmation of our findings from linear regression, but we conclude that the linear regression models are likely more reliable as they take the richness of the response variables seriously, while we had to binarize these for the logit models. At the same, we admit that our analysis leaves room for improvement. As a first step one might try to extend the linear regression models with random effects for items (Baayen et al., 2008) in order to achieve a better fit. Random effects for subjects would not make sense as each subject only saw one item. Furthermore, it would be worthwhile to conduct analyses with generalized additive models which allow the modeler to greatly relax the linearity assumption of the model. Finally, a multiclass logit analysis is called for, as well. I hope to be able to conduct those in the near future.

Interpretation and Discussion

We observe that friendliness ascriptions to the questioner decline if he asks about an inanimate object (Table 4.2). This is interesting as we observe a similar effect on the questioner’s sincerity. It seems that humans simply prefer to speak about other humans (or living entities) than dead objects. While we need supporting research here, this makes intuitive sense since (most of the time) animate beings are moving and hence generally less predictable. Therefore, the uncertainty about them is relatively greater and hence more in need of reduction. As of now, this is only an educated guess. There are likely other factors that are entangled with animacy.

As pointed out above, the unfriendly questions were those with a focus on the verb or the wh-element. It is noteworthy that these are probably the most crucial elements of a wh-question anyway. Through focusing them the question may appear as pushy. This may be a partial explanation for the observed effects. In fact, taking into account our findings about the friendliness ascribed to the answerer, it seems that whenever focus falls on a word that is very much predictable from the context but also because of the nature of the exchange (wh-questions usually ask you to make a predication about someone or something) the focus placement will be perceived as less friendly. This may indicate a connection to our conception of focus as an information spreading device. If the information is already evenly distributed and focus hence introduces a valley in the information stream (see also Figure 3.5) this may be considered an act of unfriendly behaviour. If language is optimized to evenly distribute information over the speech stream this may also be considered as an act of uncooperativity which social beings likely conceive of as unfriendly in general.

Finally, from a purely information theoretic standpoint, focus on unexpected elements may be interpreted as unfriendly because it does not reduce the listener’s uncertainty about the speaker’s intention fast enough. Hence, it is simply not very helpful for the listener. Certainly there are many other factors that figure in to social behaviour and how it is perceived. The claim put forward in this thesis is that the interpretation of most social behaviours, including language,
needs to be learned from experience. Since we view learning from experience as establishing associative links between repeatedly co-occuring events, it is only natural to view the impact of focus on friendliness and sincerity ascriptions from the same angle. In particular, this means the best humans can do is to predict future events from what they observe and evaluate their observations against what they have previously learned. This is were uncertainty reduction comes into play, again. If we are serious about information theory and the clearly-defined notions of uncertainty and predictability it provides, it is much easier for us as scientists to make the connection between different forms of behaviour. In the present case, this connection could arguably be made between the linguistic functions and social functions of focus.

In line with what we assumed when designing the experiments, our effects also show some influence of alignment. As witnessed by Table 4.1, answers that are aligned so as to mimick what the questioner said are perceived as extremely unfriendly. This goes counter our expectation which assumed alignment to be beneficial to social behaviour. How pervasive the alignment effect is remains an issue for future research.

Besides giving us some insights into the social functions of focus our data also shed some light on the definition that we initially gave for focus (def. 1). With respect to focus in questions, we stayed fairly agnostic, simply defining it by its surface form. With our results about how questions are perceived when the focus is varied, it seems that def. 1 also applies to them. The mere act of asking a question makes the occurence of wh-elements and question-final verbs highly likely. Again, I make no claim that this exhausts the function of focus in questions. However, it seems that our definition is pathing the road to broader focus definitions since it is one of the few that arguably also applies to focus in questions.
Chapter 5

Conclusion and Outlook

This thesis was written with two goals in mind. Firstly, I wanted to investigate the connection between focus and theories of language learning and processing based on information theory. Secondly, I wanted to establish that focus is not merely a linguistic phenomenon but that it also has social repercussions. In order to do so I first established an empirically founded definition of focus. While this definition is by far not exhaustive, I argued that it is currently the one that delivers most empirical adequacy. Therefore, any broader definition of focus will have to subsume the one presented here.

We have assumed throughout this thesis that language is a form of human behaviour and that, like most forms of behaviour that develop after birth, it has to be learned. This was motivated by the observation that there is no compelling evidence that would justify a special status of language amongst other human abilities. In fact, while the theory presented here can straightforwardly be grounded in widely acknowledged behavioural learning mechanisms, any theory that assumes language-specific cognitive faculties will have to make a non-obvious connection to the acquisition of these. If the language-specific faculties are assumed to be innate one would need to show when and why in evolution they should have developed. If they are assumed to be learned in early stages of life, an appropriate learning mechanism has to be formulated. As of now there is no theory about language-specific cognitive faculties that would have accomplished one or the other.

At the beginning of this thesis, the basic phonetic properties of focus marking in English and other languages such as Mandarin Chinese were established. It has been shown that phonetic focus marking is achieved through an expansion of the pitch range of the speaker’s utterance, durational lengthening on the focused word and, to a lesser extent, a rise in intensity. In addition, languages may employ post-focus compression as well as post-focus pitch lowering. Duration and post-focus compression in particular have been shown to be highly salient cues for listeners who have to identify the focus in an utterance. For most European languages, this description of focus marking is just as adequate as it is for English. It has furthermore been shown that learners can indeed discover
focus just by being exposed to utterances containing it. Besides the phonetic properties of focus, we have also discussed its known functions in great detail. It turned out that the often-proposed view of focus as a signal for alternatives is hard to justify empirically. On the other hand, there is good evidence that focus enhances memory for words so as to make them better accessible for later re-use in discourse. Also, focus shows some signs of being an attention-attracting or attention-shifting linguistic device.

Coming to the original part of this thesis, we made the connection between focus and language processing in general. Recent research has shown that language processing is highly influenced by predictability. Predictability in turn can easily be quantified through the information-theoretic notion of surprisal. I have argued from the definition of focus followed here that focus tends to fall on words that are rather unpredictable. To support this claim, I have linked a property of phonetic focus marking, namely durational lengthening, to the fact that unexpected words tend to be pronounced longer, in general. This shows that if focus is properly defined and known facts about language processing are taken into account it is relatively straightforward to arrive at accounts that at least in part explain the phenomenon in question without additional stipulations. Information theory serves as a convenient formalization tool. It greatest benefit when it comes to modeling linguistic phenomena is that its notion of information is purely quantitative and hence fully general. Notions of information that are tailored to specific aspects of communication become superfluous this way.

I also grounded the proposal about focus and predictability in a general behavioural learning model that is based on the widely accepted Rescorla-Wagner-equations. Due to its domain-generality the model fits perfectly with the view of language as behaviour. An additional benefit is that it delivers cognitively plausible explanations of frequency effects. Since information theory based on probability theory and probability distributions need to estimated from frequency in practice, we could ground the notion of predictability employed here in a domain-general learning process.

Finally, I made the connection between focus and predictability explicit and argued that it likely not only affects durational but also morphosyntactic variation. Even for languages that mark focus only morphologically, the account presented here is still viable. I then presented the results of a large-scale online study that showed that the placement of focus in questions but also focus-congruence in questions and answers has an impact on social variables such as perceived friendliness. Under the view from information theory this is easy to bring together with our findings about the connection between focus and duration: in both cases does the recipient require information to reduce his uncertainty about his interlocutor and more generally his environment. This is also the reason that Shannon (1948) advocated for why any two physical systems would need to exchange information in the first place. It seems therefore advisable to further pursue this track of research and apply concepts from information theory and behavioural learning to other phenomena of information structure. This may offer a whole new conception of information structure in
linguistics. While the very notion of information used is still very vague in the field and mostly heuristically given, the approach demonstrated here makes this notion precise. Let us take information structure seriously and turn it into a theory of how linguistic functions such as focus help humans to structure information flow! Once this is done it will also be interesting to investigate how humans interpret this information and what meanings they can extract from it. As has been shown in our experiment, these meanings by far exceed the very narrow notions of meaning that linguists have traditionally concentrated on.

Future work should and will include testing the focus-duration hypothesis of information spreading on a speech corpus. Also, a repetition of our experiment in a more controlled environment and with synthesized stimuli seems promising. The opportunities that crowd-sourcing offers for linguistic research also need to be explored further. Finally, as mentioned above, the hope is that more researchers try to apply a similar kind of theorizing to other notions in information structure. This would mean to only assume what you know but also to try to explain as much as possible with simple, domain-general mechanisms.
Bibliography


Philip Schulz. *Focus as a discourse functor*. Talk given at the Workshop on Discourse Coherence, Information Structure and Implicatures at the 25th European Summer School in Language, Logic and Information, 2013. Slides available upon request.
BIBLIOGRAPHY


