Balancing Bilingualism down the Garden Path

Eli Rugaard

Master Thesis of Linguistics
at the University of Bergen, Norway

2016

Dissertation date: May 18th, 2016

Year: 2016
Title: Balancing Bilingualism down the Garden Path
Author: Eli Rugaard

Acknowledgements

This thesis would not have been completed (or existed at all) without the help and guidance I have received. First of all, I would like to thank prof. Christer Johansson, my advisor. Thank you for encouraging me and believing in me. Thank you for trusting me by putting your name next to mine on abstracts that we have sent out. Thank you for spontaneous coffee breaks, input, and feedback throughout this process.

I am very grateful for all the participants from both the first and second round of running the experiment. This master thesis could have ended up being called ”Repetition effects: How I created, and then conducted my own experiment 122 times”, but due to so many willing participants, I could stick to the original plan.

I want to thank every teacher and professor that has taught me anything throughout this 5 year education, which is a lot of people, but especially prof. Helge Dyvik, who was very helpful in the last round of editing the stimuli for the final experiment.

The process of completing this masters thesis would have been a lot less fun without the support of my co-students. Especially, I thank Christina Stremme for setting the standard of what a masters thesis in psycholiguistics at UiB is, paving the road for Julie Sverreson and I. Julie has been of great support and help in every phase of this process. Thank you for daily support in everything from kind words, lunch breaks, written feedback and proof reading. Also, I am in debt to Victoria Troland for her encouragement and constructive criticism.

I have received two travel grants from LLE, UiB, which made conference participation a lot less damaging for my meagre student wallet. I thank the committees of PiF 2015 and Milanguage for accepting what was then a term paper, but due to helpful feedback, grew into this thesis. Also, the summer school at Radboud University in august 2015 was a great inspirational kick-start to the year of working on this thesis. I would also like to thank the friday seminar in linguistics at UiB for valuable input and feedback.

I would also like to thank kollektivet, felleschatten, Nina Stensaker and Nina Hansson, Akademisk Skrivesenter, Jakob Tønnessen, and Mia Julie Wiland. Finally, I thank my grand-mothers and my parents for non-linguistic support.

Abstract

Many factors affect how difficult a sentence is to read. Gibson (2006; 1998) proposed that people combine (a) context-dependent syntactic expectations (top–down statistical information) and (b) context-independent lexical-category frequencies of words (bottom–up statistical information) in order to resolve ambiguities in the lexical categories of words. It is well known that Garden Path (GP) sentences cause significantly longer reading time at the point of ambiguity. Similarly, Code-Switching (CS) in a sentence causes longer reading times, as words are easier to access in a context and code switching temporarily creates a change in lexical context. These phenomena can be used to explore how lexical access and sentence processing interact. I tested this with a self-paced reading (SPR) test of GP and CS sentences compared to a baseline, with an added test of balanced bilingualism in working memory. Will the code switch affect how fast the GP sentence is read and understood?

A SPR-test with sentences presented in halves (pilot study, autumn 2014) found a significant main effect of GP ($p<0.05$) and a weakly significant CS effect ($p<0.06$). The interesting finding was that when the sentence was both GP and CS there was no additive effect and the combination was faster than the sum of GP and CS ($p<0.15$). The lack of extra processing time with the added difficulty was surprising: the sentence is not built up incrementally guided by the lexical content. Therefore, CS might even prepare the reparse of the sentence.

New data from a word-by-word moving window SPR (jan 2016) shows the same result:

RT = 459.2 ms + 16.1 ms (GP) + 15.2 ms (CS) – 19.4 ms (GP + CS).

The new data collected confirm the interaction of CS and GP in a more detailed and controlled experiment. The new data collection also introduces a bilingual verbal memory task that will allow us to grade the level of bilingualism of our subjects.

Language comprehension and processing depends in part on working memory capacity (Just and Carpenter, 1992), and first and second languages may differ in how efficiently they represent words. I developed a test that may detect an imbalance in working memory capacity between L1 and L2. The hypothesis is that when we get close to the working memory limit there will be differential effects for L1 and L2, and L2 will be both slower to decide and less accurate at the limit.

Just and Carpenter (1992) allow for individual variance in verbal memory, and assert that

this is related to differences in sentence processing, for example that high-span subjects may maintain ambiguities for longer periods of time. I am interested in whether there is a differential effect between L1 and L2, with the longer-term goal to explain individual differences in syntactic processing and CS.

I used Sternberg's (1966) Memory Scanning Test (cf. (Corbin and Marquer, 2009)) as a start. Words are separated into two sets, a memory set and a search set. The memory set varied between 4, 6 and 8 words, where each word is presented for 500ms. The search set was kept constant at 4 words. The participant reads the memory set, and marks if they find a matching word in the search set. Three experimental conditions test the effect of translation A: no translations between the sets, B: one word from the memory set is translated in the search set, C: one word from the memory set is translated in the search set only when there is a match word.

There are significant reaction time effects for Memory Set at 8 words for L2, but not L1, for all conditions, which is congruent with my hypothesis that there are working memory differences between L1 and L2. I have also noted that some, but not all, participants have different distributions of correct answers between L1 and L2. This may be used as a factor that is more precise than a simple high spanner vs. low spanner test for use in experiments that involve cross-linguistic language processing.

Sammendrag

Det er mange faktorer som påvirker hvor vanskelig en setning er å lese. Gibson (2006; 1998) foreslo at menneskerr kombinerer (a) kontekstavhenging syntaktisk forventning (topp-ned statistisk informasjon) og (b) kontekstuavhengig leksikalsk kategorifrekvens av ord (bunn-opp statistisk informasjon) for å løse tvetydigheter i leksikalske kategorier av ord. To faktorer er leksikalsk tilgang og prosessering av syntaktisk tvetydighet. Det er allment kjent at villstråsetninger (engelsk garden path) gir økt lesehastighet i tvetydige punkt i setninger. Likeens gir kodeveksling økt lesehastighet fordi ord er mer tilgjengelige i kontekst, og kodeveksling fører til endring i leksikalsk kontekst. Disse fenomenene kan brukes til å utforske hvordan leksikalsk tilgjengelighet og setningsprosessering interagerer. Jeg testet dette ved en selvstyrt lesetest (engelsk self-paced reading) av villstråsetninger og kodevekslingssetninger, med en grunnlinjesetning til sammenligning, i tillegg til en test av balansert tospråklighet i arbeidsminnet. Vil kodevekslingen påvirke lesehastigheten og forståelsen av en setning?

I en lesetest med setninger presentert i halvdeler (pilotundersøkelse, høsten 2014) fant jeg en signifikant effekt av villstråsetning ($p<0.05$) og en nær-signifikant effekt av kodeveksling ($p<0.06$). Et interessant funn var at når både kodeveksling og villstråsetning er kombinert, ble det ingen sammenlagt effekt. Kombinasjonssetningene var kjappere enn summen av villstråsetninger og kodeveksling ($p<0.15$). Mangelen på ekstra prosessering ved dobbel vanskelighet var overraskende. Setningen er ikke bygget på en støtte av leksikalsk innhold. Derfor kan odeveksling forberede reparsing av setningen.

Ny data fra en lesetest med ord for ord (januar 2016) viser samme resultat:

RT = 459.2 ms + 16.1 ms (GP) + 15.2 ms (CS) – 19.4 ms (GP + CS).

Data som ble innsamlet i januar bekrefter interaksjonen av kodeveksling og villstråsetning i et mer deltaljert og kontrollert eksperiment. Ny data inkluderer og en tospråklig arbeidsminnetest som kan utvurdere deltagernes grad av tospråklighet.

Språkforståelse avhenger til dels av arbeidsminnekapasitet (Just and Carpenter, 1992), og første- og andrespråk kan avvike i hvor effektivt de representerer ord. Jeg utviklet en test som kan oppdage en ubalanse i arbeidsminnekapasitet mellom første- og andrespråk. Hypotesen er at når vi nærmer oss maks arbeidsminnekapasitet, vil det være en forskjell i effekt av første- og andrespråk. Andrespråket vil ta lenger tid og være mindre nøyaktig.

Just og Carpenter (1992) viser til individuell variasjon i verbalt minne, og påstår at dette er relatert til forskjeller i setningsprosessering, for eksempel ved at personer med høyt minnespann kan holde på tvetydige analyser i lengre perioder enn de med kortere minnespann. Jeg vil finne ut om det er en forskjell i effekt mellom første- og andrespråk, som kan relateres til individuelle variasjoner i syntaktisk prosessering og kodeveksling.

Jeg brukte Sternbergs (1966) minneskanntest (cf. (Corbin and Marquer, 2009)) som grunnlag. Minnesettene varierer mellom 4, 6 og 8 ord, hvor hvert ord blir presentert i 500 ms. Søkesettet inneholdt 4 ord. Deltagerene leser minnesettet, og svarer ja/nei på om de finner et matchende ord i søkesettet. Tre versjoner av eksperimentet tester effekten av oversettelser. Versjon A er uten oversettelser, versjon B har har ett ord fra minnesettet oversatt i søkesettet, og versjon C har ett ord fra minnesettet oversatt i søkesettet ved ja-svar, det er et matchord i søkesettet.

Det er en signifikant effekt av minnesettlengden på 8 ord for andrespråket, men ikke for førstespråket i alle versjonene, som stemmer over ens med hypotesen min om at det er forskjell i arbeidsminne mellom språkene. Jeg har også funnet at noen, men ikke all, har forskjellig distribusjon av korrekte svar i første- og andrespråket. Denne faktoren kan være mer nøyaktig enn å skille mellom høyt og lavt minnespann i eksperiment som bruker tverrlingvistisk språkprosessering.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction to the topic

We all read sentences every day; short sentences, long sentences, incomplete sentences, good sentences, or bad sentences. Even though some sentences are more difficult to process than others, we manage to process an impressive amount of sentences every day, seemingly without a great conscious effort. Most of the time, this process is quite automated and unconscious, but sometimes we face sentences that lead us astray, maybe on a trip down the Garden Path, or a sudden departure to a foreign land.

Thomas Bever (1970) wrote an article on the cognitive basis of linguistic structures that sparked a discussion on how the way we think affects the structure of our language. Bever (1970) stepped away from mapping theories that claim (in some way or another) that actual speech behavior is some regular function of the abstract linguistic structure originally isolated in linguistic investigations. Bever questioned the claim that grammar is the epicenter of all linguistic behavior, and began an exploration of approaching language as a conceptual and communicative system which recruits various kinds of human behavior, but which is not exhaustively manifested in any particular form of language behavior (Bever, 1970).

The idea that specific properties of language reflect the general cognitive laws that Bever (1970) put forward, was a contrast to the then current theory of competence vs performance dichotomy by Chomsky (Chomsky, 1965). In the light of Chomsky's (1965) theory, psycholinguists worked on revealing how cognitive constraints like working memory interact with grammar in linguistic performance (Sanz et al., 2013). When Bever (1970) proposed that some formally possible, grammatical structures never manifest themselves in natural language because children cannot understand, use or learn them, research shifted and new areas of research emerged (Sanz et al., 2013). One type of formally possible and grammatically valid, but more seldom used in natural language are Garden Path sentences. Since then, many theories have aimed to explain the processes of parsing and processing these types of sentences.

Garden path (GP) sentence is a term that was introduced by Thomas Bever in (1970), and

since then it has been well researched, both within the field of psycholinguistics and syntax. Theories of parsing of GP sentences have been proposed, such as Lyn Frazier's (1979) garden path model, or constraint-based theories, and more recently computational linguists have made models of probability of parses. Is all available information incremented at once, or is the parsing process modulated into stages?

Code-switching and bilingualism, and particularly how effortless code-switching seems to be to bilinguals, has been of interest to psycholinguists since the origin of the field in the early 20th century (Bullock and Toribio, 2009). Studying the dynamics and dimensions of bilingualism is challenging. Bilinguals can read sentences from both their languages, Grosjean's (2012) by now famous statement that "a bilingual is not two monolinguals in one head", the two languages do not operate in isolation from one another. But how do they interact? Bilingualism is often treated as a categorical variable, but the degree of bilingualism is actually a continuous variable (Luk and Bialystok, 2013). Thus it is highly interesting to investigate how bilinguals use and process language, as it may shed light on various compounds of human language processing.

Since the term GP was first introduced, there have been many theories that aim to explain the processes at work in parsing. One more recent theory of parsing involves top-down and bottom-up information. Top-down parses uses a hypothesis of general parse tree structures using syntactic expectations. Bottom-up parsing works from the low level meaning, using grammatical structure of linear input text such as lexical category and frequency (Gibson, 2006).

We know that GP induces a cognitive load caused by a reparse, that is indicated by an increase in reading time. GP breaks the syntactic expectations (top down) thus forcing a reparse, whereas CS challenges the predictability from lexical information (bottom up). Gibson (2006) proposed that people combine (a) context-dependent syntactic expectations (top–down statistical information) and (b) context-independent lexical-category frequencies of words (bottom–up statistical information) in order to resolve ambiguities in the lexical categories of words. Combining GP and CS allows a closer look at these two motivations for reparses, and how they combine. The addition of CS as a factor allows us to control and assess the impact of bottom-up processes in relation to top-down processes.

Bilinguals may be more proficient in one language than another. This can depend on the task at hand, or maybe the situation at hand. Countless studies have shown the effortlessness of switching between languages (see (Bullock and Toribio, 2009) for an overview). This effortlessness in language switching may have benefits other than being bilingual. Having two languages at hand may help the bilingual detect ambiguities that a monolingual does not reflect on. Bilinguals often show higher meta-linguistic awareness from an earlier age (Baker, 2006).

Working memory and its effect on reading has been studied by others (Just and Carpenter, 1992; Daneman and Carpenter, 1980). Individual differences in working memory has been

studied (Corbin and Marquer, 2009). Bilingualism as a non-categorical variable has been studied by others (Luk and Bialystok, 2013). Combining these factors is valuable to tell us more about the effect of bilingualism in working memory and reading.

I have not found any research that experimentally test both GP and CS simultaneously. My results can also have implications on language processing in general, and the CS makes it possible to manipulate the effect of a different encoding, although many details remain to be investigated. Combining the two may tell us something about how bilinguals process language: is reparsing affected by CS? This could be expected if the situation also involves switching between two grammars and especially integrating results from two separate grammars. I expect that both GP and CS will take longer to resolve than the baseline.

Do memory limitations play a role in reading GP sentences? Does bilingualism in short term memory matter in processing code switching sentences? I have developed a test of bilingual working memory based on Sternberg's (1966) memory scanning task. Participants will read memory sets of 4, 6 or 8 Norwegian and English words, and answer yes or no to whether they identify one word in a search set from the memory set. Two other versions of the memory test also test the effect of translations, both on matches and non-matches. This will test the effect, measured in reaction time and correct responses, of set length and language on working memory recall.

The reading task is a self-paced, non-stationary, word by word, moving window reading of sentences. Sentences come in four conditions presented in a factorial design using four blocks: a baseline, a CS, a GP, and both GP and CS in one sentence. The reading task will test the effect of CS on GP in sentences, to see if the effect is additive or interactive.

1.2 Research hypotheses

In this thesis, I will take a closer look at two well-researched linguistic phenomena, code-switching (CS) and garden paths (GP), using a well proved linguistic method, a self-paced reading test (SPR), and see how they affect the parsing process in reading. The thesis is based on a self-paced reading experiment controlled for bilingual balance through a verbal memory test. Balanced bilingualism is assessed through a bilingual verbal memory-search test loosely based on Sternberg's memory scanning test. The main experiment concerns the effect of CS in SPR of GP sentences. This can be used to address one of the key questions within bilingualism, (Wei, 2008): How is the knowledge of two languages used by the same person in bilingual interaction?

1.2.1 Self-paced Reading of Garden Paths and Code-Switching

It is predicted that there is an added cognitive load for GP and CS that is reflected in longer reading times. In the case that the sentence is both GP and CS, there are two different possi-

bilities:

H0: There is an additive effect of combining GP and CS in one sentence, the two processes are independent of each other, and both add to the processing load.

H1: The combination of GP and CS is significantly faster than predicted by an additive effect. Both processes contribute information that together makes the task easier.

1.2.2 Bilingualism in Working Memory

The working memory test will assess the effect of language in working memory, and capacity through different memory set lengths.

H0: There is no difference in the language of the match word in all set lengths of the test.

H1: There is a significant added processing load of longer memory sets and L2.

1.2.3 Correlation Effects

Also, for the memory test, we will be able to distinguish between people with high memory spans and low memory spans. This may then be reflected on the reading pace in the SPR-test. The participants who perform equally well in both English and Norwegian in the memory test, will have less of an effect of CS-ing in reading.

H0: There is no correlation between performance on SPR-test and memory test. Correct responses and reaction times in memory test do not correlate with reaction times on SPR-test.

H1: There is a correlation between performance on SPR-test and in memory test. Bilingual working memory does affect performance in ambiguity resolution of GP and reading pace of CS. Reading pace is dependent on working memory.

1.3 Definitions of terms

Here, I will give brief definitions of the key terms and concepts used for my thesis. They will be presented in depth later on.

1.3.1 Garden Path Sentences

A garden path sentence is a sentence where the first half is read unambiguously, with a preferred parse in mind, but the other half reveals that the preferred parse in the first half is not the correct one. The sentence as a whole is unambiguous, but the correct parse may be difficult to spot. The most commonly used example in English is example 1.

(1) The horse raced past the barn fell.

An Norwegian GP sentence used as stimuli in SPR-test is example 2 [1]

(2) Bonden / gir / dyr / ull / blir / klippet / av / vann.
The / farmer / gives / animal / wool / is / cut / of / water.

1.3.2 Code-Switching

Code-switching is the alternating use of two languages in the same discourse by a bilingual speaker (Bullock and Toribio, 2009). Traditionally, code-switching studies are studies of natural production of spontaneous speech, but in the experiment presented in this thesis, it is studied in the form of reading sentences, with a switch of language mid-sentence. And example of written, intra-sentential CS is presented in example 3.

(3) Det / siste / barnet / i / kindergarten / wanted / pancakes.
The / last / child / in ...

Language switching in this thesis refers to the process of switching from one language to another, although it is used differently by many others. L2 in this thesis refers to a person's second language.

1.3.3 Sentence Processing

Sentence parsing is traditionally the process of sequencing a sentence into strings. In psycholinguistics, it is also referred to as the process of comprehending language. Sentence processing is to linguistically process a sentence. As we see, these two expressions are often used interchangably. However, the term sentence processing is more widely used, and is the term I will stick to when referring to the general linguistic process.

1.3.4 Working Memory

According to (Gathercole, 2009), working memory is used in the sense of referring to the working memory system of processes involved in the temporary storage and manipulation of information, and also as a label for tasks that require the participant to store information while engaging in other cognitively demanding activities, which is what my experiment is based on. On the other side there is short-term memory, which refers to tasks that tap the storage capacities of the working memory system but impose only minimal demands on processing (Gathercole, 2009).

Working memory is a temporary store for recently activated items of information that are currently occupying consciousness and that can be manipulated and moved in and out of short-term memory (Colman, 2008)

[1] I have chosen not to gloss sentences. Word boundaries in sentences that are presented word by word are separated by "/". When the ambiguity in translation meaning of one word is essential to the GP, this is marked explicitly.

1.3.5 Bilinguals and Bilingualism

A bilingual is, as the term explains, a speaker of two languages, and a multilingual is a speaker of more than one language (Grosjean and Li, 2012). Traditionally, the separation between these two terms are not very strict. I have chosen to use the term bilingual, even though the term multilingual might be more accurate for most of my participants. I chose this because I am investigating only two languages.

Bilingual balance refers to the process of using two languages in the same way in the same context. The proficiency of the two languages the person speaks are balanced (Field, 2011).

According to Field 2011, a balanced bilingual is a person who is, in principle, highly proficient in both or all languages that he or she speaks. Implicitly, this person is fluent and accurate in either of the languages. The term ”balance” refers to that the bilingual is adept in a wide range of registers in both languages (Field, 2011). It can refer to balanced in terms of being able to use the languages in the same contexts, being equally literate in both/all languages, and reading or writing with the same ease. Other terms that are used for these concepts are ”ideal bilingual” or ”full bilingualism”. (Field, 2011). These terms have also been criticized because they could be interpreted to mean that other types of bilingualism are less than ideal.

Another term that is being used is ”ambilingual”, after the term ambidextrous - being able to use both hands equally well, the ability to use languages interchangeably (Field, 2011). The disagreement around terms is also an indicator of the problems that underly the phenomena ”balanced bilingualism”. It is very difficult to measure degrees of bilingualism, and to categorize a bilingual into types of bilingualism.

1.4 Outline of Thesis

The upcoming chapter 2 introduces CS and its three written types. I present sentence processing, and different parsing strategies, with a focus on the top-down and bottom-up strategy. Furthermore, I look at bilingualism, balanced bilingualism and how it can be studied in working memory. Then I go on to present how working memory is assessed by others.

Chapter 3 will give a more in-depth definition of GP sentences, and how they differ in Norwegian and English, I will give an explanation of the concept of CS and how I have chosen to use it. A discussion of the grammaticality in CS follows.

Chapter 4 gives an in-depth description of the methods I have chosen to research the hypotheses. I give a brief definition of reaction times and how they are used, and of the role working memory plays in reading. I give an outline of the experiments I have conducted, the methods and stimuli used, and how the experiments developed. I present the issue of data validity and the method for analysis that will be used in chapter 5.

Results are then presented in detail in chapter 5. I present the different analyses that have been conducted, and present the results of the experiments, the comprehension tests, and an analysis of correlation.

In chapter 6, I discuss the results, put them into a context, and point to what the results mean. I compare the different analyses that have been conducted and explain their meaning in relation to the theory presented in chapter 2. I also point to possibilities of further research on the topic.

Finally, based on the results and the discussion, I conclude on their meaning, and present what I have achieved through this masters thesis.

Chapter 2

Theory

Garden paths and Code-Switching are two well researched linguistic phenomena. They are interesting in their own ways, and they have been the source of many a headache in studying ambiguities or the process of switching between languages. In this chapter, I will take a look at the conditions for studying them. CSing has been studied in many ways, and I will present three types of written CS.

The experiment that this thesis is based on will test the effect of GP and CS sentences. Therefore, I will present the main theories within sentence processing, accounting for the main models of parsing strategies, in particular bottom-up and top-down parsing.

There are many studies on bilinguals, and there is an ongoing discussion of the benefits and disadvantages of bilingualism. Bilingualism in verbal working memory is an interesting factor to combine with GP and CS. Coding of languages may differ, leading to higher demands of capacity in language switching. Is sentence comprehension in one language influenced by knowledge of another?

Furthermore, I will look at working memory, and how language and bilingualism factors in, with a focus on working memory in sentence processing. Finally, I will look at findings on working memory limitations, and how it has been assessed.

2.1 Code-switching

2.1.1 Studying Code-switching

CS provides a unique window on the structural outcomes of language contact (Bullock and Toribio, 2009). It comprises a broad range of contact phenomena and is difficult to characterize definitely (Bullock and Toribio, 2009). Traditionally, social motivation for CS is usually assessed, and looking at cognitive load alone is not as common. I will step away from the most typical way of studying CS, to look at it from a more structural perspective.

CSing may be affected by a bilingual's proficiency. Studying CSing is difficult to do without saying anything about bilingualism. Combining two languages into one situation,

one sentence, or even one clause, demands an awareness of language that a monolingual may not have.

Monolinguallike control of two languages over all aspects of linguistic knowledge and use within all domains is rare, if possible at all (Grosjean and Li, 2012). Most bilinguals show disparate abilities in their component languages, for a myriad of reasons, including age of second language acquisition, the quality of linguistic input received, the language most used, and the status of the language in the community (Bullock and Toribio, 2009).

There is a debate on methodological problems regarding techniques used to study CS without compromising the phenomenon. According to Bullock (2009), studying the simultaneous activation of two languages is commonly examined through language switching tasks, to study lexical access, working memory, bilingual control or attention. Trying to induce, manipulate and replicate natural CS in a ecologically valid way is very challenging (Bullock and Toribio, 2009). See the upcoming section 4.6 for a more in-depth discussion of ecological validity.

2.1.2 Intersentential and intrasentential Code-Switching

There are two main types of CS: intersentential and intrasentential. The first refers to switching language between two sentences, or when the topic of a conversation or a text changes. This type of CS might occur when a new participant joins the discourse and there is a need to adapt, or for ease of expression in a new topic (Field, 2011).

Intrasentential code-switching is the type of code-switching used in this experiment. It refers to switching of language within a sentence boundary. In this type of code-switching, especially in spoken situations, it is assumed that both languages are activated (Field, 2011). This type of switching is not necessarily due to change in the speech situation, and may be frowned upon or considered bad style. This view may come from thinking that the languages are not spoken correctly because they are mixed, and may not be proper due to community standards or norms. Examples of this type of CS is Spanglish (Spanish and English) or Denglish (Deutsch (German) and English). The term for these types of ”broken” language combinations is macaronic language.

2.1.3 Macaronic Language

Macaronic language is to use a mixture of languages in one text (Beatie, 1967). It has often been used in poetry for humorous effect (SNL, 2009). Particularly bilingual puns such as example 4 are frequently used.

(4) We take care of your boss.
”boss” in Norwegian can mean garbage.
Seen on a garbage truck in Bergen.

Throughout history, latin has often been mixed with another language in poetry for humorous effect, for example by adding latin endings to a word, examples are pig latin, or the play ”Erasmus Montanus” by Ludvig Holberg. Macaronic language has also been used in hymns, without the intention of humor as an effect, and can also be used for lyrical effect. The genre of macaronic poetry origins from Italy in the 15th century, from medievel preaches in latin, where the intention was not to be funny, and it was then first parodied in a poem about macaroni, hence the name (SNL, 2009).

Macaronic language also has an effect other than humorous, it shows that bilinguals use a meta-understanding of language for effect. This is related to the intrasentential CS in the sense that the CS is not random, but is used to trigger meta-associations across languages and stratified language use.

Metalinguistic awareness is the awareness of how language works, and the ability to reflect and be conscious of the nature of language (Field, 2011). It involves treating language as an object of conscious thought, as opposed to using language unconsciously to comprehend and generate utterances.

Baker (2006) showed in a study that bilingual children appeared to have greater metalinguistic awareness and a more analytical view of language compared to monolinguals, and Baker suggested that the ability to control two languages enabled the bilingual children to perform better on tasks such as counting words in a sentence.

Why is CS and metalinguistic awareness interesting to combine with GP? Just as monolinguals possess intuitions about what constitutes well-formed utterances in their native language, bilinguals have the capacity to differentiate ill-formed from grammatical patterns of CS (Gullberg et al., 2009). The metalinguistic awareness that CS or macaronic language brings may make it easier to read an ambiguous sentence.

2.1.4 Code-switching and Lexical Access

The bilingual lexicon is the mental store of vocabulary items for the bilingual’s language (Field, 2011). Field defines to ”know a word” in the mental lexicon as: a) all the forms a particular word may take, b) its grammatical function, c) its core meaning and d) all the possible meanings it can have according to the speaker’s knowledge (Field, 2011). There is a dispute on how these words are stored for a bilingual. There is an ongoing discussion of how the lexicon of one language influences another, or if the bilingual has two independent lexicons, or if the words are all unified in one word bank (Weinreich, 1953; Potter et al., 1984).

Lexical access refers to the mental recognition and retrieval of lexical items from one’s mental lexicon (Field, 2011). It involves locating the particular word or word form intended for a specific meaning, and vice versa, hearing a word and recognizing its meaning. The storage of words is referred to as the mental lexicon, which is the total knowledge of words that a person has.

There is a large amount of linguistic research on the structural patterns of bilingual code-switching, see (Bullock and Toribio, 2009) for an overview. Of interest is the specific point of CS. It is clear that CS takes place at specific points in the utterance that is well formed and conforming to the grammatical constraints of the languages involved (Wei, 2008). CS is normally considered evidence of shared storage access of both lexicons simultaneously (Gullberg et al., 2009).

2.2 Sentence Processing

Language processing refers to how the brain deciphers and responds to linguistic data (Field, 2011). It also entails memory, such as declarative memory or working memory. Processing of sentences includes taking in the information and structuring it to make sense. Ambiguous sentences adds an extra demand to sentences, but the underlying processes are hard to pin down.

On-line methods are important in studying how the human language processing system accesses and makes use of different kinds of linguistic information, such as syntactic or semantic/lexical information. (Kaiser, 2013) Structural syntactic ambiguity originates from the fact that certain constituents may contract several different grammatical relationships (Kempen, 1996). Podesva (2013) gives an example.

(5) The *witness* examined by the lawyer turned out to be unreliable.

(6) The *evidence* examined by the lawyer turned out to be unreliable.

The example 5 is ambiguous, but example 6 is not. Two very alike sentences give different meanings, different structures, and thus different processing through changing one word. The problem with these sentences are that they can be read as a whole, without being aware of ambiguity. Using GP sentences, that are unambiguous as a whole, give a valuable insight into processing of sentences.

Research in both syntax and psycholinguistics has offered models of how sentence parsing is achieved. There are two main theories: interactive accounts, such as (Gibson, 2006, 1998), and modular accounts, such as (Rayner and Frazier, 1987; Frazier, 1979). Interactive accounts uses all relevant information immediately, but in modular accounts some information can be used immediately, but some cannot (van Gompel and Pickering, 2009). Much of sentence processing literature implicitly assumes a specialized syntactic processor (Niv, 1993). Lately, some have stepped away from giving an explicit model of the processes, but uses probability to explain preferred parses.

Bever (1970) influenced the view on incrementality in sentence processing. Since then, it has been assumed that both syntactic parsing and semantic interpretation are highly incremental. Readers update syntactic and semantic interpretation on a word-by-word basis as they read (Staub, 2015).

"Sentence processing research has shown that parsing is largely incremental, i.e. language comprehenders incorporate each word into the preceding syntactic structure as they encounter it; they do not delay syntactic structure building until, for instance, the end of the sentence or phrase" p. 289 (van Gompel and Pickering, 2009).

Another point of discussion within sentence parsing is whether it is serial or parallel. This differs from incrementality. Incrementality states that the information is incremented as soon as possible. But does it construct just one syntactic interpretation or are multiple parses kept at hand? This is especially important in ambiguous sentences.

A serial model claims that one analysis is kept in mind, and as information is added incrementally, that analysis proves to be incorrect, a reanalysis occurs and a new analysis arises (Staub, 2015). An example of this is Frazier's (1979; 1987) garden path model.

In contrast, a parallel model claims that the parsing process may maintain several possible analyses at the same time, with degrees of activation (Staub, 2015). In ambiguities, more than one parse will be kept in mind until it is disambiguated. MacDonald, Just and Carpenter (1992) gathered data that suggest that the alternative lexical frames are both activated and maintained while the ambiguous region is being processed. This is in line with the constraint-based theory.

There is also a combination theory of the two, the unrestricted race model, where at the point of ambiguity, multiple analyses race to be constructed (van Gompel et al., 2001).

The two main theories, the GP model and constraint-based theories, will explain the GP effect differently. According to a serial parsing theory, the excess cognitive load is due to a reparse, but according to constraint-based theories, it is due to the strain of keeping alternative parses in mind.

2.2.1 Parsing Strategies

Parsing stategies involve the syntactic processes of structuring a sentence. In the following, I will focus on theories that have been presented as parsing strategies for GPs.

A lot of research has focused on determining whether all available information is used during the earliest stages of processing, for example (van Gompel et al., 2001; Rayner and Frazier, 1987; Frazier, 1979; MacDonald et al., 1992).

In Bever's (1970) by now legendary paper "The Cognitive Basis for Linguistic Structures", he argues that specific properties of language reflect cognitive laws. One of the main questions Bever raised was how can a person arrive at internal linguistic knowledge from external input sequences?

Since then, a series of processing theories and parsing preferences have been presented. Of imporatance has been the involvement of semantics in sentence processing at early stages, as well as other language-independent cognitive factors (Sanz et al., 2013).

There are two main types of parsing models. *Modular models* assume that the mind consists of modules that perform very specific processes These processes are informationally

Input -> Lexical Processor -> Category -> Syntactic Parser -> Syntactic Structure -> Thematic Interpreter -> Sentence Meaning

Figure 2.1: Lyn Frazier's Garden Path Model of Syntactic Parsing

encapsulated: they use only information represented within this module (van Gompel and Pickering, 2009). *Interactive accounts* assume that the processor immediately draws upon all possible sources of information during sentence processing, including semantics, discourse context, and information about the frequency of syntactic structures (van Gompel and Pickering, 2009).

Another pioneer within studies of GP is Lyn Frazier (Frazier, 1979; Rayner and Frazier, 1987). Frazier proposed a two-stage model of syntactic parsing. The first stage analyzes words to determine what categories they belong to. When categories have been identified, the parser builds a syntactic structure. The main input of information for this step is word category. The second stage computes meaning from applying semantic rules to the structured input. The model is serial and modular. The process is visualized in figure 2.2.1:

Frazier follows the modal principle of two-stage model with two heuristics: minimal attachment and late closure. Minimal attachment means that the parser builds the simplest structure possible, i.e. the structure with the fewest nodes. Late closure attaches as many words or phrases to the ongoing clause. The process is serial. Late closure claims to not postulate unnecessary structure. If possible, continue to work on the same phrase or clause as long as possible (Traxler, 2012). Minimal attachment claims that if more than one structure is licensed and consistent with the input, build the structure with the fewest nodes.

The most prominent alternative theory to the garden path model is the constraint-based parsing model. There are two main differences. As opposed to building one structure at a time, this parser pursues multiple structures simultaneously (Traxler, 2012). This is explained by a parallel process. Constraint-based parsers represent different aspects of sentences, with syntactic structures and activation patterns (MacDonald et al., 1992). Another difference to the garden path model is that where the garden path model relies on word category information, the constraint-based model draw on a wider variety of cues to decide what structures to build (Traxler, 2012). This is considered a one-stage model, as opposed to Lyn Frazier's two-stage model. It is parallel, and not separated into stages.

Ford, Bresnan and Kaplan (1982) argue that aside from purely structural ambiguity resolution criteria, the parsing processor is also sensitive to the "strength" of association between certain words like verbs and the nouns they take as arguments, e. g. transitivity. Many current theories explain the "strength" with predictability. Predictability can be separated into two main fields: syntactic predictability - the amount of structural possibilities, or word frequency. Van Gompel (2001) states that the initial analysis of a syntactic ambiguity is affected both by individual differences and by syntactic and non-syntactic characteristics.

A model by a computational linguist Niv (1993) claims that the syntactic processor is

the simplest imaginable. It merely represents the syntactic analysis of the input. Niv claims that resolution of ambiguity is performed by the interpreter (Niv, 1993). He describes four criteria:

1. Plausibility of the message carried by the analysis
2. Quality of fit of this message into the current discourse
3. Felicity of the constructions used in the utterance to express the message
4. The relative frequency of use of a certain construction or lexical item (Niv, 1993)

Another more recent theory of parsing is (Gibson, 2006). Gibson (2006) proposes that the human sentence processing mechanism pursues all grammatically available analyses in parallel as it processes the string, discarding those analyses which are 'too costly' — that is, when the cost of one analysis, A, exceeds that of another analysis, B, by more than P Processing Load Units, A is discarded, necessitating conscious effort to reconstruct should it be subsequently necessary.

In 1998, Gibson presented a theory of the relationship between the sentence processing mechanism and the available computational resources. The theory was named the Syntactic Prediction Locality Theory (SPLT) and consists of two components, the integration cost and the memory cost (Gibson, 1998). The memory cost in this theory is the cost associated with keeping obligatory syntactic requirements in mind. Memory cost is hypothesized to be quantified in terms of the number of syntactic categories that are necessary to complete the current input string as a grammatical sentence (Gibson, 1998). In line with previous studies on working memory, he claims that both memory cost and integration cost are heavily influenced by locality and distance. He defines it as:

1. Locality: The longer a predicted category must be kept in memory before the prediction is satisfied, the greater is the cost for maintaining that prediction.
2. Distance: the greater the distance between an incoming word and the most local head or dependent to which it attaches, the greater the integration cost.

2.2.2 Bottom-up and top-down

Gibson (2006) investigates how people resolve syntactic category ambiguities when comprehending sentences though SPR-experiments. He proposes that people use two different types of information to resolve syntactic ambiguity:
a) context-dependent syntactic expectations - top-down statistical information
b) context-independent lexical-category frequencies of words - bottom up-statistical information (Gibson, 2006)

The concept of top-down and bottom-up approaches to language are based on the observation that language consists of a hierarchy of forms (Field, 2011).

Bottom-up parsing refers to parsing from the lowest level upwards. It focuses on form and the construction of language knowledge from the basic building blocks of language upwards to the higher level of meaning in discourse (Field, 2011). Bottom-up parsing is based on the idea that information starts at the lowest level of information blocks. Bottom-up is not only used within parsing strategies, but is also used as a theory for language teaching which focuses on form and the basic building blocks of language.

On the other side of bottom-up parsing we have top-down parsing. In top-down parsing, meaning is the driving form of parsing or acquisition. As opposed to bottom-up information flow, top-down parsing does not emphasize form (Field, 2011).

Gibson (2006) presents the results of a previous study by Tabor et al called ”Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing”. This study provided evidence that people are sensitive to the syntactic context in resolving lexical category ambiguities. From this research, Gibson (2006) proposes a context-independent lexical category-frequency hypothesis: People are sensitive to the lexical category-frequency distributions of each word, independent of context.

Another of Gibson’s (2006) hypotheses is one that rules out theories that claim many trees are kept in mind. He suggest error-detection-based reanalysis: the human sentence processor attempts reanalysis when the lexical entry that has been selected cannot be incorporated into the current phrase structure tree.

A summary of the propositions Gibson (2006) gives:
a) the processor tracks frequencies of syntactic categories for words independent of syntactic context
b) the processor tracks syntactic expectations in different syntactic environments
c) syntactic expectations are smoothed upward to a minimum level for very infrequent syntactic expectations
d) the lexical and syntactic weights are multiplied together in order to get a relative weight, which serves as an estimate of people’s preferences.

2.2.3 Semantics in Sentence Processing

Syntactic parsing is the process of the analysis of grammatical structure of each sentence, and the word’s place in this structure (Staub, 2015). Semantic interpretation is the process of combining meaning of individual words and phrases to construct a representation of the sentence’s overall meaning (Staub, 2015).

Online studies of sentence processing brings up the question of modularity, a key theme in psycholinguistics: does the language processing system use both syntactic and semantic

cues (as well as other cues) when parsing a sentence (an interactive system) or is the system modular - in particular, do early stages of processing only make use of syntactic information? (Kaiser, 2013).

Is syntactic processing separable from other processes such as semantic processes? Even if the processes were known, there are different theories about how they might affect reading times. The process could be additive and serial, they might interact, or they might operate in parallel (Gernsbacher, 1994). The main difference between syntactic and lexical information in sentence processing is that the human syntactic processor avoids keeping multiple analyses, but it seems to consider alternatives when processing lexical items (Kempen, 1996).

Semantic information often provides strongly constraining information for syntactic analysis, so an important question has been whether this information is used immediately to guide sentence processing (van Gompel and Pickering, 2009). According to modular accounts, there are stages of processing that delay the information. According to constraint-based models, semantic information should have an immediate effect on sentence processing, whereas according to modular models (p. 293)" (van Gompel and Pickering, 2009). Changing the language of the sentence that is read forces the reader to activate a new semantic context.

2.2.4 Working Memory in Sentence Processing

The possibility that the working memory system supports the processing of language for meaning has been discussed within the field of linguistics. Often, working memory is assessed using complex memory span tasks such as Daneman and Carpenter's (1980) reading task, which tests processing of a sentence and storage of words.

Other studies have found a correlation between individual differences in memory span and aspescts of sentence processing, see (Gathercole, 2009) for an overview.

Just and Carpenter (1992) claim that individual differences in people's verbal memory lead to individual differences in sentence processing. They claim that high-span subjects maintain ambiguities for longer periods of time, where the nature of a persons language comprehension depends on his or her working memory capacity. In the same paper, Just and Carpenter also propose a shared resources account of working memory in which all linguistic processes draw upon the same limited pool of working memory resources.

Caplan and Waters (1999) comment on Just and Carpenter, and suggest a different solution in which working memory resources are dedicated to obligatory and automatic linguistic processes, and sentence processing is different from the resources used for more strategic and controlled processes such as those used in Daneman's (1980) reading span test.

MacDonald et. al. (1992) found that there was a correlation between working memory and ambiguity resolution, where high-spanners were more able to keep parallel representations of potential sentence representations, whereas low-spanners did not.

So to relate this to my own experiment, Just and Carpenter (1992) say there is a correlation between reading and working memory, Caplan and Waters (1999) say they are different

processes and therefore not comparable. MacDonald et. al. (1992) claim that the correlation between reading and working memory depend on how good their working memory is. According to Just and Carpenter, there will be a correlation between my memory test and SPR test. According to Caplan and Waters, they are independent of each others. Lastly, according to MacDonald et al, there will be a correlation on a per-participant basis.

2.3 Bilingualism

2.3.1 Code-Switching and Bilingualism

”Bilingual” is a cover term that encompasses speakers who fall along a ”bilingual range”, a continuum of linguistic abilities and communicative strategies (Luk and Bialystok, 2013). The words ”bilingual” and ”bilingualism” have many different meanings depending on the context in which they are used (Grosjean and Li, 2012). Within the field of psycholinguistics, a general definition is that it is the use of two or more languages or dialects in everyday life (Grosjean and Li, 2012).

A common misconception that surrounds bilingualism is that the bilingual has equal and perfect knowledge of their two or more languages, that they have no accent in any of their languages, that they acquired their languages in childhood, that they are competent translators, and so on (Grosjean and Li, 2012). I will look further into the first part of this statement, the knowledge of the two (or more) languages a bilingual masters.

One very important characteristic of the bilingual is their ability to move between different languages (Wei, 2008). The bilingual can use their languages one at a time like a monolingual, or mix languages in the same sequence (intersentential CS), clause (intrasentential CS) or even word (macaronic langugage).

Bilinguals can be described in terms of language use and language fluency. There might not always be a correlation between the order of which the language has been acquired and language use, or language use and language skill. A bilingual usually (but not always) has a dominant language, but the dominant language may change over time due to overall use of language, domains covered by a language or a combination (Grosjean and Li, 2012). Therefore, it might not always be safe to assume that a person’s mother tongue (or first learned language) is automatically their dominant language (Grosjean and Li, 2012).

2.3.2 Balanced Bilingualism

I am interested in bilingual balance as measured by working memory for each lexicon or language. A balanced bilingual does not have to be measured in explicit knowledge of two languages, nor does it have to mean that the language first acquired is the language that is strongest. An index of bilingual balance can say something about the bilingual’s implicit knowledge of languages tested in working memory, cirrekated ti CS in reading fluency.

One study by Marian and Neisser (2000) showed in an experimental study that events are better recalled if the language used to recall them is the language in which the event took place. They called it ”language-dependent recall”. This can be related to the effect of translations in the memory test study.

Most of research on the cognitive organization and representation of bilingual knowledge can be traced back to Weinreich’s (1953) ”Languages in Contact: Findings and Problems”. Weinreich distinguished three types of bilinguals, where he focused on the relationship between the linguistic sign (signifier) and semantic content (signified). The three types are

a) coordinate: the individual combines a signifier from each language with a separate unit of signified
b) compound: the individual identifies two signifiers but regards them as single compound
c) subordinate: the individual learns a new language with the help of a previously acquired one
(Weinreich, 1953)

A balanced bilingual will not have a stronger relation between a signifier and a signified in each of his/her languages. The compound will be a balanced bilingual in meaning relations, and may not have a preference in meaning activations. The subordinate may in theory exclude balanced bilingualism because it implies a hierarchy of language knowledge.

Luk and Bialystok (2013) claim that bilingualism is not a categorical variable, and that interaction between language usage and proficiency has to be studied. They make the important claim that it is not sufficient to classify a participant as monolingual or bilingual, but individual variation in aspects of bilingual experience needs to be established (Luk and Bialystok, 2013). This approach assumes that bilingualism is best described as a multidimensional construct rather than a categorical variable. The aim of their study was to quantify bilingualism using multifactor statistical analysis. They tested this through formal tests of English proficiency and a self-report questionnaire (Luk and Bialystok, 2013).

2.3.3 Language and Bilingualism in Working Memory

Language skills include the passive abilities of listening and reading, and the active abilities of speaking and writing (Field, 2011). How does working memory affect the language skill reading? Memory capacity and keeping words in short-term memory may have an effect on reading of sentences. Working memory also combines task-solving. Maybe working memory can be related to the ability to see alternative parses of a sentence?

We have now discussed parsing strategies, and working memory is an important factor in parsing. What role does working memory play in sentence processing? Working memory is not a single store, but a memory system comprising separable interacting components. Working memory may affect the reading pace, especially when a sentence is ambiguous,

and also comprehension of sentences may be influenced by working memory.

The question of whether memory limitations are responsible for another form of processing difficulty, namely garden path sentences is much more controversial (Niv, 1993). Therefore we will see the SPR experiment in the light of the test of balanced bilingualism, to see if there is a correlation between performance in the memory test compared to the reading test.

As we saw in in the introduction there is a distinction between working memory and short term memory. They are related, but traditionally they are separated in how they are used in literature. Working memory has two main definitions; it refers to the working memory system of processes involved in the temporary storage and manipulation of information (Baddeley and Hitch, 1974), as a label for tasks that require participants to store information while engaging in other cognitively demanding activities (Gathercole, 2009). Short-term memory on the other hand, refers to tasks that tap into the storage capacities of the working memory systems, but only require minimal processing (Gathercole, 2009).

”When people’s working memory capacity is exceeded, because either high storage or processing demands are very high, this should result in either a processing slow down or a failure to maintain linguistic information in memory” (van Gompel and Pickering, 2009) p 297.

Studies in bilingual sentence processing have focused on phenomena related to how semantic or syntactic representations are built (Hernandez et al., 2009). Processing occurs on both the semantic and the syntactic level. One of the most frequently asked questions within the field of bilingual language processing is the question of whether a bilingual stores both sets of words for each language in the same or in separate lexica. Many models set out to explain this, but it is not within the scope of this thesis to present them.

2.4 Assessing Working Memory

In the previous section we looked at claims correlating working memory and sentence processing. Before we take a look at the methods I have used, let’s take a look at ways to assess working memory.

Baddeley and Hitch (1974) proposed a now famous model of working memory that describes the processes of working memory in three components: central executive, phonological loop, visiospatial sketchpad, and they later (2000) added the episodic buffer. They were the first to introduce the concept of working memory, and have given a basis for studying it further. It is a temporary store for recently activated items of information that are currently occupying consciousness and that can be manipulated and moved in and out of short-term memory (Colman, 2008)

Just and Carpenter (1992) assessed working memory through Daneman and Carpenter’s (1980) reading span task. In this task, participants read series of unrelated sentences out

lout, and the task is to remember the final word of each sentence in a series. The amount of sentences increased until a participant could no longer correctly recall the final word (Daneman and Carpenter, 1980). This is an off-line method (off-line and on-line methods will be discussed in chapter 4).

Caplan and Waters (1999) claim there are different linguistic processes of working memory. One process involves the on-line, unconscious, psycholinguistic processes of comprehension, and the other involves controlled verbally mediated task. According to this division, Daneman and Carpenter's (1980) task only involves the controlled working memory. Caplan and Waters (1999) also consider the relationship between individual differences in working memory and individual differences in the efficiency of syntactic processing. In their study from 1995, they had participants match sentences to pictures in a no interference condition and in two concurrent verbal load conditions: while retaining a random sequence of digits equal to their span and equal to one less than their span (study together with Rochon, cited in (Caplan and Waters, 1999)).

In 2002, MacDonald and Christensen came with a criticism of Just and Carpenter's (1992) and Caplan and Waters' (1999) claims. MacDonald and Christensen (2002) claim that there is an unnatural split between linguistic and non-linguistic working memory, and they claim that individual differences in comprehension do not stem from variations in a separate working memory capacity; instead they emerge from an interaction of biological factors and language experience (MacDonald and Christensen, 2002). In this article, they raise the very important question of what working memory tasks really measure.

Both Just and Carpenter (1992) and Waters and Caplan (1999) agree that the reading span task provides a measure of some kind of working memory capacity, but Just and Carpenter see this capacity as central to language comprehension whereas Waters and Caplan do not.

Both Just and Carpenter's (1992) and Caplan and Waters' (1999) experiments are off-line assessing only one language.

2.4.1 Miller's Magical Number Seven

George Miller (1956) has done pioneering work on defining working memory. His theory is that people can keep an average of seven items in mind at a time, with an individual variance of plus or minus two. In his article, he presents a study on pitch by Pollack, where performance is nearly perfect up to five or six different stimuli but declines as the number of different stimuli is increased (Miller, 1956).

His studies started with experiments on what was then called absolute judgement. This tested how accurately people can assign numbers to the magnitudes of various aspects of a stimulus. Nowadays, these types of experiments are called information capacity experiments.

There is a clear and definite limit to the accuracy with which we can identify absolutely the magnitude of a unidimensional stimulus variable. Miller (1956) maintains that for unidimensional judgments, this span is usually somewhere in the neighborhood of seven. But

what about testing two langugages, does it count as two dimensions? Is there an extra load of activating L2?

Miller's claim has been tested in many ways, using for example numbers, words or letters as stimuli. Miller (1956) concludes that the span of absolute judgment and the span of immediate memory impose severe limitations on the amount of information that people are able to receive, process, and remember. He claims that the stimulus is organized into dimensions and chunks to challenge this span. Miller (1956) points out that linguistic recoding is essential to the thought process and may be the source of individual differences. Miller's (1956) method provided a new way of quantitative measures of memory span. His theory is a point of reference for stimulus materials in measuring performance of participants.

Miller (1956) does not attempt to give an answer to why the number is seven.

2.4.2 Sternbergs Memory Scanning Task

Sternberg (1966) introduced a task to measure memory scanning effects. The task is to read a memory set, a list of items such as numbers or words. The participant will then be presented with an item that may or may not have been present in the set, and is asked to respond "yes" or "no" accordingly (Sternberg, 1966). The reaction time is recorded. Sternberg (1966) found that RT varies with the length of the memory set. This is known as the Serial Exhaustive Search Theory (Sternberg, 1969).

In his original memory scanning task, he showed participants a sequence of one to six digits and asked them first to decide whether a given test digit was in the sequence, and then to recall the digits in the order in which they had been presented (Sternberg, 1966). Since then, many studies that are based on Sternberg's task have omitted the second part of the study (Corbin and Marquer, 2009).

His famous task yielded two main findings:

1. Mean response time increases linearly with sequence length.

2. The slope of the line is the same for both yes and no reponses (but no is slower) (Sternberg, 1966, 1969)

Since then, his task has been replicated, duplicated, changed and edited. One replication by Corbin and Marquer (2009) found that in very short memory sets (1-3) items, different strategies were used than on longer sets. They also point out that Sternberg's task is very complex because it comprises six sequence lengths and two types of responses, which increases the number of possibilities of procedures (Corbin and Marquer, 2009).

Chapter 3

Material

Linguistic research often suffers from definitions and theories being based on the English language. Therefore I will take a closer look at how GP occurs in Norwegian and in English, and see how the English definition of a GP works for Norwegian.

Before we can take a look at the experiment I have conducted, we must look at the research it builds on, and the research that is the basis of my experiment. I will give a definition of Garden Paths that will be used to define the sentences used in the experiment.

CS is traditionally researched in spontaneous speech, but I have approached it through taking a look at different forms of written CS. Therefore, I will look at the grammaticality aspect of combining two languages in one sentence.

3.1 Types of Garden Paths

3.1.1 Garden Paths in Norwegian and English

To be able to assess GP in an experiment, we need to have a clear definition of what a GP is, and what makes a GP in Norwegian, and to see if the rules and research on GPs in English is valid for Norwegian too [1]. Constructions of GPs are language specific phenomena, and an English GP sentence does not necessarily translate as a GP into Norwegian. There has been plenty of research done on English GPs (see (Sanz et al., 2013) for an overview), but not as much has been written about GPs in Norwegian. Before we go on to conduct an experiment combining GP and CS, we have to look at GPs in both English and in Norwegian.

The term Garden Path was coined by Thomas Bever in 1970. A garden path sentence is a sentence that is temporarily ambiguous because it contains a word or a clause that appears to be compatible with more than one syntactic structure (Bever, 1970). When reading the sentence, the preferred parse is the wrong parse, thus leading the reader down the garden path. The correct but dispreferred parse is often overlooked because it may be difficult to see

[1] I have further investigated this in a term paper for the course DASP307 at University of Bergen spring 2015

other alternatives than the preferred parse.

GP sentences are unambiguous, but on the path to a complete sentence the reader will be predictably led to start on a wrong path, which is unambiguously detected when the sentence cannot be completed with that first choice. The wrong first choice is predictable as the choice most people would make in sentences of the same structure. This is in contrast to most ambiguous sentences, where the ambiguity is often not detected.

”In many instances the semantic relations and unique lexical classifications in English can themselves determine the segmentation that creates GP” (Bever, 1970). GP is not a given structure that can be defined by language-specific regulations. There are sentences with the same structure as GP, where the initial wrong choice is either not made, or much easier to resolve, possibly for semantic reasons.

(7) The horse raced past the barn fell.

(8) The car raced past the barn crashed.

The structures of the two sentences are the same. This shows that a GP is not only a sentence structure, but is also guided by lexical information. For example animacy makes it more probable to read ”the car” as passive, but ”the horse” is more animate, and thus the predictions change.

According to Traxler (2012), GPs lead to less accurate comprehension, and a larger cognitive load in the brain while reading. GPs are mostly a written phenomenon. When reading aloud or speaking, prosody will tell us which words that are grouped together. Sentence processing can be helped by words or affixes which explicitly mark the syntactic structure, but these are often left out in English without making the utterance ungrammatical, or creating a GP (Warren, 2013).

I used a definition of GPs from (Jurafsky and Martin, 2009) that is a summary of Bever’s (1970) famous article ”The Cognitive Basis of Linguistic Structure”. This definition gives three properties to define GPs. The definition is based on English as an example, but we will now see if the definition is also valid for Norwegian, and if the definition draws a line between GPs and other lexically or structurally ambigous sentences. One question that must be answered before I conduct the experiment is if GP is its own cross-linguistic syntactic phenomenon, or if there are shades of gray between GPs and other ambiguous sentences. These are the three traits that Bever (1970) gives to define a GP:

1. They are temporarily ambiguous: The sentence is unambiguous, but its initial portion is ambiguous.

2. One or two or more parses in the initial portion is somehow preferable to the human parsing mechanism.

3. But the dispreferred parse is the correct one for the sentence. (Jurafsky and Martin, 2009)

The first trait is the one that most clearly separates GP from other lexically and structurally ambiguous sentences. GPs have a first part that is ambiguous, but it is in the second part that we become aware of the ambiguity of the first part. An important trait that separates GP from other ambiguous sentences is that we are always mislead down the GP, but in other ambiguous sentences one can read the whole sentence without being aware that there is another interpretation of it.

Jurafsky and Martin (2009) explain the ”somehow” of the second trait with probability. We prefer the parse that is most probable from either syntax or word frequency. Other possible influences are according to Jurafsky and Martin memory span, thematic structure and discourse limitations (Jurafsky and Martin, 2009). The experiment that will be conducted, will combine a test of memory span with a test of reading pace of GP. The ”somehow” in the definition gives a big span that allows it to be adapted to the language in mind. What does the ”somehow” entail in Norwegian and English?

The third trait of the definition states that the dispreferred parse is the correct one. The first part will lead us astray, and then in the second part, we find that the unlikely parse is indeed the correct one. The most significant difference between a GP and another ambiguous sentence is that in the GP, one really trips. A sentence such as sentence 9 can be read and parsed while being unaware of its ambiguity.

(9) The hunter killed the poacher with the rifle.

A garden path is an open definition of a type of ambiguity, and not one type of syntactic structure. Therefore, it is not always possible to translate a GP. The definition leaves out some other types of syntactic ambiguities.

Cross-linguistic differences in relative clause attachment present a problem for the garden-path theory, because late closure predicts a universal preference for low attachment (van Gompel and Pickering, 2009).

Since language has the capacity to generate an infinite number of sentences, it is very difficult, if not impossible, to give a complete list of structures or words that allow for a GP in one language. Still, there are traits that are easily detectable when comparing two languages.

It can sometimes be hard to distinguish a GP from a reduced relative. There is a big overlap between GP in Norwegian and English, even though they separate on some points. The definition Bever gives is specific enough to separate GP from other lexically and structurally ambiguous sentences, but still allows for some room for language specific definitions.

3.1.2 Is a GP really a GP?

We have established that the definition of a GP excludes many other forms of ambiguous sentences, but we have also seen that the definition is very open for interpretation, and leaves room for many different types of constructions within the definition.

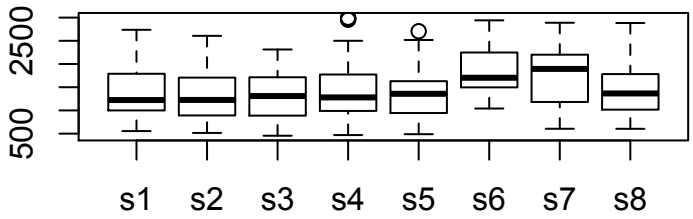

Figure 3.1: Pilot Study: Box plot of RT for GP
The x-axis shows reading pace in ms. Reaction times under 400 ms and over 3000 ms are excluded. The different GP sentences are represented by s1, s2...

Jurafsky and Martin (2009) also introduce the term mini garden paths . An example of a mini-garden path is example 10.

(10) The students forgot the solution was in the back of the book.

These mini-GPs make it difficult to draw a clear line between GPs and other lexically or structurally ambiguous sentences. The sentence is not a complete GP, because even though the clause boundaries in the first half are confusing, it does not give a significant change in meaning. Using the term ”mini-GP” makes it more difficult to separate between GPs and other lexically or strucually ambiguous sentences.

We have seen that GPs are more complex than lexically or structurally ambiguous sentences, but there is also a big variety within the group of GPs.

Now I will get a little ahead of myself by presenting some results of the pilot study. A box plot 3.1 of the RTs of the sentences used for the pilot study shows variation within the group of GPs (methods will be presented in the next chapter). The box plot 3.1 shows the variation of reaction times within the group of GP sentences.

The reaction times varied within the group of GPs. As the box plot 3.1 shows, sentence 6 (given as example 11) and sentence 7 (given as example 12) took longer to process.

(11) Det minste barnet kunne gjøre | i buksa av og til.
(The smallest child | The least the child) could | go in his pants sometimes.

(12) Kvinnen som har kastet som målte | lengst av alle vant.
The woman who has (thrown | the throw) measuring | longest of all won.

(13) Per gir fisken middag | blir laget av god mat.
Per gives the fish dinner | is made of good food.

Sentence 12 is a reduced relative sentence within a reduced relative sentence. NP ”Kvinnen som har kastet som målte lengst av alle” VP ”vant”.

In setence 11 it’s harder to point at something that separates this sentence from the others. ”Det minste X kunne gjøre” is a phrase that occurs often, so parsing it differently from the idiomatic meaningm might be more difficult.

On the other hand, sentence 13 (sentence 8 of the box plot) did not take longer than the rest of the GP sentences, even though it contains the greedy object, Per gir OBJ ”fisken middag blir laget av” mat.

3.2 Grammaticality in Code-Switching

Van Kesteren et al (2012) raised the simple but complex question of ”How do bilinguals realize to which of their languages a word belongs?” They pointed out that there has not been much research done on this. There are models of lexical access in bilinguals, aiming to explain the processes of reading, listening and speaking and language membership, but still many researchers claim that it is not necessary to explicitly represent this information as much is explained by context (van Kesteren et al., 2012). But what happens when this context changes?

Norwegian and English are similar languages in many ways. Still, there are structural differences that make GP sentences different in the two languages. Also, there are structural rules that need to be obeyed when forming a grammatical sentence in CSing.

First of all, what is a grammatical CS sentence? This is a tricky question to answer. Moyer (2008) states that CS needs to conform to the grammatical constraints of both languages. This can mean two things. It needs to obey the grammar of the two languages, or it needs to obey the syntactic structure of both languages.

Structural differences between verb positioning in Norwegian and English can make CS sentences ungrammatical.

(14) While the woman was dressing the child played outdoors.

(15) While the woman was dressing barnet lekte utendørs.

(16) While the woman was dressing the salad barnet lekte utendørs.

(17) While the woman was dressing the salad lekte barnet utendørs.

The first example 14 is a non-CS English GP. The second example 15 is ungrammatical by Norwegian rules for subordinate clauses, where the verb comes first, as opposed to the rule

for main clauses, where Norwegian uses V2. This example 15 does not fulfill the structures for a grammatical sentence for both languages, and may therefore be seen as ungrammatical.

The same goes for example 16, it does not follow structural rules for Norwegian subordinate clauses.

Example 17 may be deemed grammatical because at the point of CS, not only do the words switch language, but so do the structure rules. The use of ”while” in these sentences, lead to a subordinate clause, and with a CS, the rules for a grammatical sentence changes.

A lexical decision task study on French-English bilinguals by Bauvillain and Grainger (1987) found that RTs were slower if words were presented in a mixed language list than in a monolingual list. Participants took even longer if the word’s ortography did not exclude one of the languages.

Language membership information from both the lexical and the sublexical level can potentially facilitate word recognition (van Kesteren et al., 2012). But when a word can be from either of the languages, it makes processing more difficult. In Norwegian, there are three added letters of the alphabet: æ, ø, å. They will immediately reveal the language origin of the word, facilitating word recognition.

3.2.1 Translational homographs at the point of CS

A translational homograph is a homograph that occurs across languages, such as ”is” being an English word that in Norwegian means ”ice cream”. It is not an aim of this experiment to study translational homographs. Still, the sentences have to be carefully constructed as to not contain cognates at the point where CS occurs. Some translational homograph between English and Norwegian may make a sentence ungrammatical.

Two example sentences from the experimental stimuli where it may have been a problem are given in 18 and 19. In example 18, the sentence can be interpreted as two Norwegian nouns following each other, making the sentence ungrammatical.

(18) Per gir fisken middag is made of food.
Per gives the fish dinner (ice cream|is) made of food. Stimulus sentence:
Per gir fisken middag er made of food.

This problem can be avoided by pushing the point of CS forward.

(19) Det minste barnet kunne do was clean up.
The (least|smallest) the child could (toilet|do) was clean up.
Stimulus sentence:
Det minste barnet kunne gjøre was clean up.

In example 19, the problem lies in the translated meaning of English ”do” meaning ”toilet” in Norwegian. Language-specific properties of the input string may be especially relevant to the word recognition process in bilinguals, because they could be used to discriminate

word candidates from target and non-target language (van Kesteren et al., 2012). This is an important point in grammaticality judgement of these types of sentences, and should to the greatest extent possible be avoided.

One example from the stimulus where I failed to detect a translational homograph at the point of CS is the example 20

(20) Forfatteren / mente / at / everyone / should / read / the / book.
The writer / gave opinion / at=that ...

At can be both Norwegian and English. It is a highly frequent word in both languages. This may pose a problem, or it will be solved in the same way a GP is read. In the GP, the reader is first unaware of the ambiguity of the rest of the sentence, which may be assumed at the point of CS, where the reader is not likely to activate the CS spontaneously. If there is such an effect, it may be an argument for a constraint-based theory on the lexical level, and is another example of how CS affects the syntactic structure.

3.2.2 Syntax across languages

Norwegian and English share many traits when it comes to syntactic rules. One difference is the inversion of verb order in Norwegian sub-clauses.

(21) Han skal ikke spise opp maten hennes.
He shall not eat up her food.

(22) Han sa at han ikke skal spise opp maten hennes.
He said that he not shall eat up her food.

This may pose a problem for grammaticality judgement in CS. A study by Eve Higby et al. (2015) found that the concept of shared syntax for bilinguals extends to language-specific constructions as well. We have seen that there are many aspects of two languages that need to be taken into consideration while judging grammaticality in CS.

Chapter 4

Method

This chapter will give an outline of the experiments that have been conducted: the pilot study, experiment 1 and experiment 2. Through all three versions of the experiment, one part of the experiments is the self-paced reading test, although it goes through some major and minor changes. The pilot study has a more exhaustive comprehension test, and in experiment 1 and 2 there is the added memory test. I have developed a test of balanced bilingualism in short-term memory, using translation priming to see how it affects the process.

I presented Gibson's (2006) theory on syntactic ambiguity resolving, and his proposition that people use syntactic (top-down), and lexical (bottom-up) information to resolve ambiguity. Gibson (2006) used three SPR experiments that were conducted involving the ambiguous word «that» in different syntactic environments in order to test his theory of bottom up versus top down information in processing. His data supports the top-down/bottom-up approach in which the relative frequencies of lexical entries for a word are tabulated independent of context.

So I will take syntactic ambiguity in the form of GP, and lexical category change in the form of CS, to see how these two affect the reading pace of a sentence, and how they act together in resolving ambiguity.

This chapter introduces the pilot study from another course, the first experiment that was conducted, and the second and final version of it. I will not go as into depth of the pilot study, as it is only used as a basis for the preceding experiments . I will not go as deep into the details of the first study either, as the data and results have not been used, but I will go through the adaptions made, and the lessons learned before I present the method of experiment 2 in full detail.

4.1 Reaction Time and Self-paced Reading

There are many ways to study reading and bilingualism within the field of psycholinguistics, but measuring reaction times is the technique that has been around for the longest. In 1868, Fransiscus Donders suggested that one could infer the time taken up by a particular hypo-

thetical mental stage by subjecting a participant to two procedures that differ only in whether that stage is used (Luce, 1991).

Reaction time studies have been used to study many different factors of language and language processing. It is often assumed that longer reaction times are associated with increased processing load and processing difficulty (Podesva and Sharma, 2013). Reaction time studies have been criticized because they can teach us something about overall organization, but very little about the details of processing (Luce, 1991). Self-paced reading (SPR) methods are based on the assumption that a subject reads a passage at a pace that matches the internal comprehension process (Luce, 1991).

SPR studies have lately been overshadowed by eye-tracking studies, but they are still in use today because they require little equipment, and are less expensive than fMRI, ERP or eye-tracking equipment. SPR is also criticized for being less fine-grained than eye-tracking, but it has still been proven to capture subtle aspects of processing (Kaiser, 2013). Other criticism towards reaction time studies point out that reading times does not reveal the source of the changes in processing, but this is also true for eye-tracking. Even if these processes were known, there are different theories about how they affect reading times. The process could be serial or sequential, interacting or parallel (cf: (Luce, 1991) for detail).

Within the field of response time studies, there are three types to be distinguished (Luce, 1991). *Reaction time* is the time it takes to react to a stimuli or a task. *Reading time* is a more fluent task that does not require a decision, it simply measures that stimuli is taken in, this is an example of the type of RT used in the SPR-test. *Response time* is the time it takes to respond between two or more answers to a question or a task, such as in the memory test. For simplicity, I will refer to them all as RT, but they indicate different underlying differences in the RT of the reading test and the RT of the memory test.

On-line methods investigate real-time aspects of language processing. These methods play an important role in psycholinguistic research, because many of the processes underlying human language processing are very rapid (on the order of milliseconds), transient, and not accessible to introspection. On-line methods allow us to gain insights into transient effects that are often not explicitly ”noticed” by language users, and also make it possible to learn about the time-course of both language production and comprehension. Because many psycholinguistic theories make explicit claims about the relative timing and relations between different aspects of language processing, on-line methods often play a crucial role in allowing us to compare competing theories. (Kaiser, 2013).

Within reaction times, three different types have been defined (Luce, 1991). *Simple reaction times* are reaction times obtained in experiments where participants respond to stimuli such as light, sound and so on (Baayen and Milin, 2010). The SPR-test is an example of this, where the participant will read word by word. *Recognition reaction times* are also called go/no-go task, and is an experiment where the participants respond to some of the stimuli, but ignore distraction stimuli. The last type is *choice reaction times*, where the participants

select a response from a set of possible responses, such as in the memory span test, where the participant will have to decide if they recognize a word from the memory set. Within the definition of choice reaction times, there is the more specified *discrimination reaction times*, which are obtained when subjects have to compare simultaneously presented data and respond typically yes or no (Baayen and Milin, 2010).

4.2 Working Memory in Reading

Working memory may affect how a person processes a sentence. Just and Carpenter (1992) claim that only high-span individuals have sufficient working memory capacity to use contextual information to disambiguate syntactic ambiguities. Working memory is crucial in the parsing process of ambiguous sentences. If one is not able to keep all the words one has read ready in memory, one will not be capable of the extra processing that an ambiguous sentence requires.

There are many ways to assess working memory capacity. A common method is to assess reading span tests (Daneman and Carpenter, 1980). Their test is to read aloud each of a sequence of sentences, and recall the terminal word. This way of testing combines both processing and storage. They found that individual differences in complex memory span relates to several aspects of sentence processing. King and Just (1991) compared reading times to reading comprehension and found that low-span participants were particularly slow in reading the critical part of the sentence in ambiguous sentences.

MacDonald et al. (1992) found that individual differences in memory span were also linked to the resolution of syntactic ambiguities. They found that high-span participants appeared to construct parallel representations of potential interpretations prior to the resolution point in the sentence, but low-span participants did not do this. Just and Carpenter (1992) claim that only high-span individuals have sufficient working memory capacity to use contextual information to disambiguate syntactic ambiguities.

Including two languages into one test of working memory will assess bilingual balance. It is important to specify that testing bilingual balance is not testing explicit knowledge of a person's languages. It is not measuring use of each of the languages. It is simply measuring the accessability of resources in working memory through bilingual balance. I aim to test how the resources of the working memory are used when two languages are at hand. This is measured both in reaction time per language, and correct responses per language.

4.3 The Pilot Study

4.3.1 Experimental Procedure

In this study, participants were all students between the ages of 19-30, and were gathered voluntarily from the University campus. Participants were not compensated with money for participating, but were offered candy. In the pilot study, we sampled 62 subjects, but only 36 performed to criteria. Some were excluded because they did not show that they had read and understood the sentences, and most because they gave no response, maxing out the time slot of 3000 ms, on a significant number of the sentences, thus giving too few data points to analyze. The eligible participants all responded within the allowed time, with a lower bound for completing the task at 400 ms. Reading times that are less than 400 ms implies that the key press was done before the stimuli was read. Reading time slower than 3000ms are most likely due to other processes than reading, for example distraction [1].

4.3.2 Experimental Design

The experiment was designed on a macbook pro using SuperLab 4.5 to create the stimuli data. The experiment was presented on a pink flat screen TV that was connected to the macbook pro. The participant responses were collected with a Cedrus RB-530 response pad. The response pad offers 2-3 ms accuracy, as opposed to up to 30 ms resolution accuracy from a normal keyboard. The data files were handled using Cedrus Data Viewer 2.0. The experiment was conducted in a soundproof lab, including a chair, a desk, and the equipment used to present the experiment in the room. The participants conducted the experiment alone in the room, but were told that they could leave the room at any time if they wished.

The reading pace is measured in reaction time, being exposed to the first half of a sentence, and then having to press a key to see the second part, and similarly after reading the second part.

(23) Mannen som hører stemmer | spiller fint saksofon.
The man who hears voices | plays saxophone well.

(24) Mannen som hører stemmer | en fin saksofon
The man who hears tunes | a nice saxophone.

(25) Mannen som hører stemmer | plays saxophone well.
The man who hears voices | plays saxophone well.

[1] This study was conducted as part of the course LING306 in the first semester of my masters degree at University of Bergen. The experiment was planned and conducted with two other students. The results of the experiments were presented as poster presentations at *Psycholinguistics in Flandern* and *Milanguage* in may 2015 together with Christer Johansson.

(26) Mannen som hører stemmer | a nice saxophone.
The man who hears (voices|tunes) | a nice saxophone.

The experiment was split into four blocks of eight sentences. Every sentence comes in a baseline version (23) , a garden path version (24), a code switched version (25) , and a code switched garden path version (26). Each participant will only see one of the four versions above. A garden path and a non-garden path version of the same sentence was thus never displayed to the same participant. This was done to exclude repetition as a factor. Each participant did one of the four blocks, and each block had an equal amount of participants. Each block had four training sentences and four lead in sentences, and eight test sentences. With four blocks, that gave us 32 non-repeated versions. Each block has approximately the same amount of words (about 30). Sentences were randomized.

The sentences were displayed in halves, with a focus point:

*
and masks
#####
preceding them.

When using self-paced reading, it is common to include comprehension questions to ensure that participants are paying attention to the experimental stimuli (Kaiser, 2013). Comprehension was tested by control questions after each sentence. Information about whether they responded right or wrong were also collected. The subjects were to choose the sentence with the same meaning as the test sentence out of four possibilities. Feedback on whether the response was correct or not was not given.

(27) Velg setningen som tilsvarer setningen du nettopp leste:
Saksofonen blir spilt av den hørende mannen.
Saksofonen blir stemt av den hørende mannen.
Saksofonen blir spilt av mannen som hører stemmer.
Saksofonen blir stemt av mannen som hører stemmer.

Translation:
Choose the sentence corresponding to the sentence you just read:
The saxophone is played by the hearing man.
The saxophone is tuned by the hearing man.
The saxophone is tuned by the man hearing voices.
The saxophone is played by the man who hears voices.

The subjects had to show variation in their responses to check that they were not just pressing the keys without reading, i.e. the correct response was placed in all four positions ABCD.

The results were analyzed using a linear regression model with random intercepts for subject and sentence.

The subjects were also tested for comprehension by choosing the sentence with the same meaning as the test sentence out of four possibilities. We chose to use all 61 subjects and 8 tested question responses per subject (N=488).

4.3.3 Problems with pilot study

Having no similar previous experiments to base ours on, we faced some problems upon conducting our experiment. One problem was the instructions. We had conducted a preliminary pilot test on one person, and had changed the instructions to fit the feedback, but we soon realized that the instructions given on the screen were insufficient. To compensate for this, we gave more careful instructions in person to each participant before he or she started the experiment, instructing them as they conducted the first example sentence. This proved somewhat successful. Still, out of 62 test participants, only 36 performed to task. Upon completing the experiment, the participants were asked to give comments.

The comments varied, some commented on the language switching, many commented on the sentences being «strange», and most were curious as to what was really going on.

The pilot experiment yielded interesting results, but it also came with some problems. One unexpected problem was that not all subjects could perform the task to criteria. We had planned the reading pace to allow at most 3000 ms for reading each part of the sentence. This is enough time to read, but it is not enough time to stop and think. Reading with an average reading pace of 500-1000 ms per section of the sentence compared to reading at a pace of 3000+ ms per section is two very different reading strategies that are not comparable.

One explanation to why there were so varying results in reading pace is due to the difficulty of the comprehension questions, and this likely has biased subjects to read the sentences very carefully. The problem could be fixed in many ways, one way is to give more elaborate instructions and a longer training session with feedback. Another obvious idea is to separate the check questions from the reading task. I will give an elaborate explanation of what I chose to do in the next experiment in the upcoming section.

We did not test the participants memory span, nor did we test English proficiency. We know that all participants were Norwegians under the age of 30 studying at the University of Bergen, so we know that they have completed a minimum of 10 years of English in school. Also, in almost all (if not all) studies at University of Bergen, some of the curriculum is in English, so we can assume that they are comfortable reading English.

4.3.4 Improvements

We discovered that subjects may solve the task differently. It could be that better instructions, longer training blocks with feedback and separating the reading task from the comprehension task may give higher task compliance. It could also be that the differences between fast and slow readers are correlated with their status on bilingualism. It could also be that some subjects read more fluently than others, possibly due to factors such as short-term memory or working memory. Since task compliance in this experiment was fairly low, there are likely more factors to consider than we planned for.

When the sentences were presented in halves, sentence clauses were often separated in a way that favored the non-GP parsing, and thus pushed the reader down the garden path instead of just leading. Reading word-by-word will not give an advantage to either of them. As stated by Warren (2013), the clause is the basic unit of analysis in language comprehension, and processing is concentrated at clause boundaries. Therefore, separating sentences in the middle of clauses will then make the reading less natural, and possibly favor GPs.

4.4 Experiment 1

This experiment was conducted using the same equipment as in Experiment 2, I will give more detail about method in the upcoming section of Experiment 2.

4.4.1 Experimental Procedure

48 participants conducted the experiment. There were two versions of the memory test and four different SPR-blocks. 24 participants conducted each of the memory tests, with 12 in each SPR-block. 6 participants conducted memory test A and 6 memory test B for each block.

This experiment was conducted on students in the age span 19-34. There were 48 participants, of which 5 were excluded because of reading/writing-impairments, one was excluded because he/she did not respond to the memory test, and one was excluded because he/she performed the experiment with a baby on his/her lap. All 5 participants diagnosed with reading or writing impairments still performed the experiments to task, but performed noticeably slower and with more errors.

In experiment 1, the participants were asked, in Norwegian, a few questions about their use of English before starting the experiment. The questions were:

When did you start learning English?
How often do you use English? (use = hear, speak or read)
When you watch English tv/series/movies, do you use subtitles?

The answers of these questions were used as a control for English usage and as an indicator of whether there is a difference in how the participants use English, with the possibility of correlating it to the results of the memory span test.

The answers were quite unanimous. All started learning English between the ages of 6-8 and all used English daily or almost daily. On the last question, there was a little more variation, but most said that they do not actively choose it. These answers indicate that the participants are a quite homogeneous group, and that it also shows that these students may use English more often than the general population of Norway. Testing on a homogeneous group has its benefits in that there may be more unanimous results, and that results indicating a difference may be more robust because the participants have the same background.

4.4.2 Stimuli

The stimuli used in this experiment was unbalanced. In the memory test, there were three set lengths of the memory set. In the short (4) search set, 3 of the match words were Norwegian and 1 English, and in the long (8) search set, 3 were English, and only 1 Norwegian. This makes it harder to score well on English compared to Norwegian, and is not balanced. So the results of this experiment will not be used, neither the data of the SPR, because it will no longer be possible to compare performance in SPR between high spanners and low spanners. Still there were things to be noted. The results indicate that the experiments were answered as expected, with good task compliance, which is important information since this is a new experiment. Also, people with reading/writing-disorders were easy to spot as outliers, and will have to be excluded.

In the SPR-test, there was feedback that was worth noting for the next round. The first word of the sentence always takes longer to read, and there is a greater variation between participants here than in other word positions of the sentence. The SPR-test was controlled for word amount in each block (100 +/- 2), but another point that might simplify the statistical process and make the sentences more comparable would be to format the sentences into containing the same amount of words. The stimuli was then modified for Experiment 2, where sentences are formatted into containing 7 or 8 words each, making them more comparable.

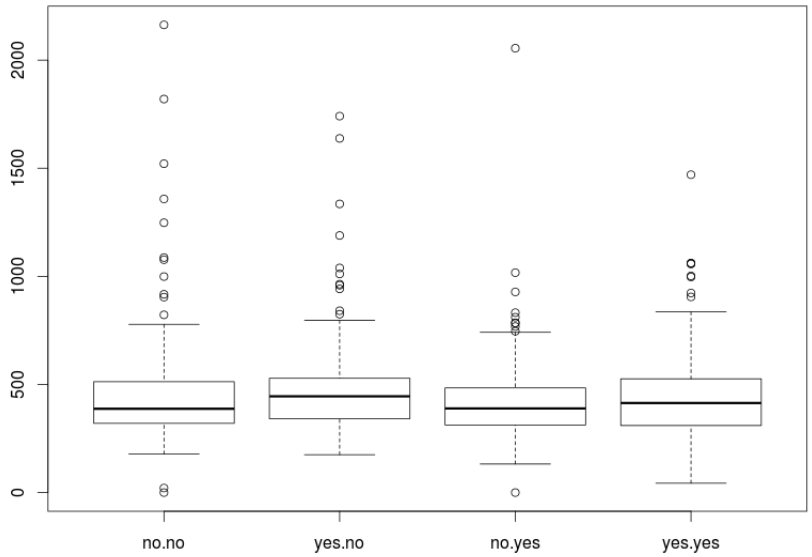

Figure 4.1: Experiment 1: Box plot of RT for Block 2
no.no = Baseline, yes.no = GP, no.yes = CS, yes.yes = GP+CS

The figure 4.1 shows average RT for block 2. As we see, there is very little variation. This is coincidentally because there were fewer words in the GP sentences. There is an increase in RT on a per word basis, but since there are more words that are baseline, there is not much effect visible here. This box plot 4.1 also shows that participants did solve the task as expected, there are very few outliers above 1500, and the RTs are quite uniform, indicating that they have solve the tasked as expected.

4.5 Experiment 2

4.5.1 Experimental Equipment

The equipment was the same for experiment 1 and 2. The stimuli was presented using SuperLab 5.0 on an Apple mini mac. The response times were gathered from a Cedrus RB-530 response pad. The RB Series response pads offer 2-3 millisecond reaction time resolution. By contrast, USB keyboards' resolution is around 10 milliseconds, and PCs using PS2 keyboards have a resolution between 20 and 35 milliseconds (Ced, 2016).

The data was presented on a Samsung model U24E590D screen, an upgrade from the screen used in the pilot study. The experiment was conducted in a soundproof room (not the same one as in the pilot study), with a desk and a chair facing away from the door. Beside the door there is a small window where the experimenter can observe the participant in action. Data was then handled using Cedrus Data viewer 2.0 and Cedrus File Merger. All

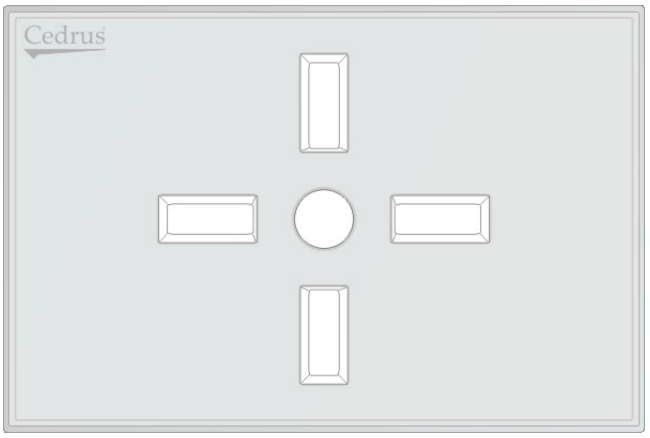

Figure 4.2: Cedrus RB-530
Response pad used in experiment had horizontal keys marked green (right) and red (left) to clearly indicate correct and wrong response keys

experiments were conduced with myself as the experimenter.

4.5.2 Experimental design

The main difference between the pilot study and experiment 1 and 2 is that the comprehension part is very modified. It has changed from being a research factor to test comprehension, to be a control factor to confirm that participants are indeed reading the sentences. The type of comprehension test used influences the reader's goal and reading time. When verbatim recall as opposed to question answering is required, the reader slows down and may change his or her reading pattern (Gernsbacher, 1994).

(28) Velg setningen som tilsvarer setningen du nettopp leste:
Saksofonen blir spilt av den hørende mannen.
Saksofonen blir stemt av den hørende mannen.
Saksofonen blir spilt av mannen som hører stemmer.
Saksofonen blir stemt av mannen som hører stemmer.

Choose the sentence corresponding to the sentence you just read:
The saxophone is played by the hearing man.
The saxophone is tuned by the hearing man.
The saxophone is tuned by the man hearing voices.
The saxophone is played by the man who hears voices.

(29) Inneholdt setningen du nettopp leste ordet "sko"?
Did the sentence you just read contain the word "shoe"?

As we can see from example 28 and 29, the changes in comprehension questions were quite drastic. In the pilot study, comprehension was a factor we were interested to see the results of, in experiment 1 and 2, the comprehension questions are just used as controls to test if the participant has solved the task as wanted, meaning that they read each word of the sentence.

Changes have also been made to the self paced reading test. Instead of presenting the sentences in halves, the sentences are now presented word by word in a non-stationary, single word moving window, with masks before and after the word presented. Word by word moving window is the most frequently used method of measuring RT in SPR-studies (Podesva and Sharma, 2013). In this method, only one word is presented at a time, and dashes represent the other words in the sentence. The word presented is concealed by a button press, and the previous word is replaced with a dash, and the next word appears. Dash length correlates with word length:

___ __ hører ___ __ __ ____

With each key press, the previous word is masked, and a new one appear, and the words appear to march across the screen at the pace set by the reader (Gernsbacher, 1994). This method allows researchers to record how much time the participant spends on one word before moving on to the next one, which can shed light on which points in a sentence are associated with increased processing load or processing difficulty (Podesva and Sharma, 2013). Also, when presenting more than one word at a time, the participant may begin processing a text segment prior to fully fixating on it (Luce, 1991), which is why it's ideal to present only one word at a time. The sentences are presented in the middle of the screen, with a fixation point * preceding it.

Instructions have been revised, and the training session changed. In the training session, there is a loop effect if the participant reads a word in < 200 ms or > 1000 ms, forcing the participant into a reading pace that is greater than 200 ms per word, and slower than 1000 ms. This is done so that all participants are forced to read at a similar pace, instead of at their own leisure. Normally, it takes a person 150+ ms from the brain reads the stimuli to it sends signals to the hand to respond in the response test to press to see the next word, and then some time for processing before they press, so for them to actually read the word, at least 200 ms is needed (Baayen, 2009).

The experiment now contains 4 lead-in sentences, and 12 test sentences divided into four sections.There are also four question sequences, where the sentence before the question and the first sentence after the question are excluded from the 12 test sentences due to possible ring effects of the question in reading the sentence after the question. Of course, a ring effect before the question is presented is impossible because the participant does not know that a question is coming at that point, but it is excluded nonetheless. So all together, the participant

will read 4 lead ins, 12 test sentences, 8 sentences surrounding the questions, and 4 questions. The sequence is randomized (except for the lead ins), so that it is not predictable when a question will come. The lead ins are not randomized and are the same across all blocks. An example of a question sequence is given in example 30

(30) Anne / slår / hunden / katten / plaget.
Inneholdt setningen du nettopp leste ordet ”hunden”?
Sykkelen / i / skuret / is / newer / than / mine.

Translation:
Anne / hits / the dog / the cat/ aggravated.
Did the sentence you just read contain the word ”the dog”?
The bike / in / the shed / is / newer / than / mine.

The question sequences include all types of test sentences. A complete list of stimuli is given i Appendix B.

Here is an example of how it may be for two fictive participants. First, the participant reads 4 lead-in sentences. They are looped, forcing the participant to read the words in less than a second, training them to read in a pace more similar to how one normally reads. This is done to teach the participant to solve the task as wanted. It may be that one of our two fictive participants gets it right away, pressing through the sentences to read. The other participant however, struggles at first. The participant reads halfway through the sentence, hesitates, and is thus forced to read the sentence again. The participant learns to speed up, and is allowed to continue. The participants do not know that only the first 4 sentences are looped.

After the lead ins, the experiment may look different for the two participants. One may get a question sequence (sentence, question, sentence) right away, then read 5 sentences, then two question sequences following each other. The other may have them more evenly distributed. It is randomized.

Changes in the experiment design from Experiment 1 to 2 is that one extra comprehension question is added. In Experiment 1, questions come after every fourth sentence, but in Experiment 2, the sequence is completely randomized. To compensate for this, participants are informed that the questions are randomly placed after some sentences, and that they might feel a bit sudden. Other changes from experiment 1 is of course that the stimuli has become balanced, and the addition of a new condition for the memory test. Sentences also underwent some minor changes, the most important change is that all sentences were edited to contain 7 or 8 words, to make them more comparable.

Similar for both Experiment 1 and Experiment 2 is that they share the same distribution of sentences across blocks. A factorial design (Kaiser, 2013) is used for the SPR-test to avoid repetition effect. 12 sentences, each in four versions: GP, CS, baseline, and GP+CS = 48

sentences distributed into four blocks. For example, sentence 12 will look like this in the different blocks:

(31) Block 1, CS:
Bonden / gir / dyrene / food / and / dry / hay.
The farmer / gives / animals ...

(32) Block 2, GPCS:
Bonden / gir / dyrene / wool / is / cut / from / water.
The farmer / gives / animals ...

(33) Block 3, BL:
Bonden / gir / dyrene / mat / og / tørt / høy.
The farmer / gives / animals / food / and / dry / hay.

(34) Block 4, GP:
Bonden / gir / dyrene / ull / blir / klippet / av / vann.
The farmer / gives / animals / wool / is / cut / from / water.

Here we see the four versions a sentence can occur in. In block 1, the sentence 31 has 7 words, in block 2 sentence 32 has 8 words, in block 3 sentence 33 has 7 words, and finally in block 4 sentence 34 has 8 words. Unfortunately, it is impossible to manipulate the sentences to all maintain 7 (or 8) words while still fulfilling the demands of being a grammatically correct sentence, a GP, or a CS. The sentences are randomized to control for training effect (e.g. learning how to solve the task) or effect of tiredness or boredom.

4.5.3 Memory Test

After conducting the pilot study, we saw there was a need for a memory test to be able to control for balanced bilingualism in correlation with SPR of GP and CS. Testing the readers' memory span might also tell us something about whether there is a correlation between high spanner - low RT on GP+CS, or whether the effect GP+CS is additive for low spanners? All in all, this memory test will give a clearer indication of individual performance.

In order to account for both the language level and verbal memory span, I developed a test inspired by Sternberg's Memory Scanning Test (1966). The idea behind Sternberg's memory scanning task is that a subject is to judge (yes/no) whether a test symbol is contained in a short memorized sequence of symbols (Sternberg, 1966). It is then expected that their mean reaction-time increases linearly with the length of the sequence. The linearity and slope of the function imply the existence of an internal serial-comparison process whose average rate is between 25 and 30 symbols per second (Sternberg, 1966). Traditionally, the test uses letters as input. Also, it is the identity of the symbols in the series, but not their order, that is relevant to the binary response.

My test varies from Sternberg's original test in some ways. Instead of using letters as input, full words are used. In the original test, 24 practices + 144 tests are used, where each string is presented for 1.2 seconds. There are 3 lead in rounds, and 24 test rounds where RT and correct response is collected, and two languages for input words.

The aim of the test is to see how the participants react to L1 words compared to L2. This will give an indication of bilingual balance, as working memory might differ from L1 to L2. One study by Marian and Neisser (2000) found that recalling was dependent on the language the memory was "stored in". This may indicate no effect of translations in the memory test.

This test comes in three conditions and uses memory lists of words with half of the words in English and half in Norwegian. The words are separated into two sets, a memory set and a search set. The participant first reads the memory set, and will then perform a task after having read the search set. It will look something like this:

Memory set:

skrue ___ __ ___ __ __ ____

Search set:

BJEFF LUE PEBBLE BRANCH

The pacing of the memory set is fixed at 500 ms per word, and will automatically display the next word of the set, no button pressing involved. The memory set words are presented all at once, in capital letters to control for form. The participant will then have to decide whether one of the words in the memory set are reoccuring in the search set. Half of the words have a match in the search set, so the correct response is yes (there is a match) and half do not have a match, so the correct response is no (there is no match).

The baseline (version A) is the task as explained above. The main purpose of this version is to serve as a basis for comparison for the following two versions. Since this is a new way of testing balanced bilingualism in working memory, it is important to have a simple baseline for comparison. The memory test is used as a dependent variable of the SPR.

(35) A: 4 related
button / peace / tau / kinn
DRUE PEACE SQUARE BØLGE

Then there are two other versions. In the second version (B), there is a distractor effect of translations of words of the memory set in the search set 36. All of the words have a translation in the search set. This serves to be able to test for the translation effect alone (when there is no match).

(36) B: 4 unrelated
kjeks / nail / blad / sheep
BIL FRAME PANTS SAU

In the third version (C), only the yes a with one match in the search list. The other condition has in addition a translated version of a word from the memory list in the search set, otherwise the same task. This will test the effect of a translation on a match, as opposed to only a match in version A.

(37) C: 4 related
bille / fork / smør / oat
BILLE SEED LANGUAGE HAVRE

In example 37, the first four words are the memory set, and the last four capitalized words are the search set (where one searches for the match).

We controlled for position by placing the target word in all positions of the search list A, B, C or D for the four positions in the search set, one match word in each position ABCD for each of the set lengths, adding up to three A, three B, three C and three D per participant. Expected yes or no answers were balanced by having four yes-related, or no-unrelated for each set length.

Memory sets have three levels: short (4 words), mid (6) and long (8), with four lines for each of the lengths, allowing match words to be in all positions for all lengths, and having an equal amount of no match-lists, adding up to 8 lines x 3 different set lengths = 24 data points of reaction time and correct response.

Memory sets are presented in small letters, and the search sets have capital letters to control for form. This forces the participant to actually register the meaning of the word, not only scan for similarity of form to recognize a word. The task is to decide if any word appears in a search list of four different words, half Norwegian and half English.

Trials are randomized, so that the difficulty is random, and not increasing, to avoid learning effects of effects of tiredness.

4.5.4 Experimental Procedure

Before the experiment was started, the participants were asked if they had ever been diagnosed with dyslexia or reading disorders, whether they had impaired vision, whether they were epileptic, whether Norwegian was their first language, and their age. If the participant had been diagnosed with reading disorders, they were still allowed to complete the experiment, but were excluded later on in the process. Participation was completely voluntary, and the participants were informed before the experiment started and before the second part started that they could leave the room if they wished to. They were also informed that the experimenter was waiting outside the window, and that they could wave to the experimenter

if they had questions after reading the instructions, before starting the experiment, or at any other time. The participants were allowed to read the instructions at their own pace.

The participants were told that if they hesitated or stopped in their reading, in the beginning the sentences would start over. They were also informed that there were questions following only some of the sentences, and that they are randomly placed, so they might take the participant a little by surprise. Both parts of the experiment takes approximately 10 minutes altogether to complete, where the memory test takes a little more time than the SPR-test. The participants did not know what the purpose of the experiment was. The participants did not receive any compensation, but were offered a small snack.

The extent to which the results can be generalized depends on the selection of participants (Lanza, 2008). For this study, participants were recruited through student Facebook groups for students in Bergen, also posters were placed on campuses in town. But most of the participants were recruited from asking in person on the campus of faculty of Humanities at University of Bergen. In this process, it is important to try to avoid selection bias, as people have a tendency to be drawn to friendly faces or people of a certain type (Wray and Bloomer, 2012). So in recruiting from a cafeteria, I mapped out a route, and asked at the tables according to this route, to avoid asking only one type of people (i.e. the friendly faces).

When generalizing over results, it is important to consider the representativeness of this group. As mentioned, the majority of participants for this study were students in their 20s, so the results may not be representative of the population of Norway in general, but the experiment can of course be extended to include a more representative group in the future. But for reaction time studies, having people in the same age group might create clearer data, as other studies have shown that older people react slower in reaction time studies (Baayen and Milin, 2010).

4.5.5 File treatment

Before the data can be analyzed, it needs to be formatted and processed in such a way that patterns of regularity can be observed, measured, and presented (Moyer, 2008).

File formatting was carried out using SuperLab filemerger, LibreOffice Calc, gedit text editor and R (RStudio Team, 2015).

A t-test and standard deviation calculated in R is used to select the 99 percentile, so reaction times under 200 and over 1201 are excluded in the SPR-test. Memory files were tagged manually in excel for set length, language of the match word, and translation direction (for B and C). RTs under 350 were removed.

SPR-files were tagged for point of GP and CS. Sentence length could have been controlled for when creating the stimuli sentences, but it is hard to compromise on word length while creating a GP or a grammatically correct CS according to English phrase structure. Also, excluding the first word is a good idea, as there is a larger span between RT in the

first word. Sentences were then formatted so that in sentences with 7 words, the first word was removed, and in sentences with 8 words, the first two words were removed, so that all sentences have 6 data points (given that none have been excluded as outliers in previous file handling rounds). Sentence 10 in Block 3 has only 6 words, so it remains complete or will be excluded.

This first analysis will give an indication of overall effect, and then we can take a closer look at effect at the point of CS or GP (or both), tagging the files for -2, -1, 1 and 2 around the point of GP, CS, or the point where GP would have occurred in the baseline sentences. As mentioned above, in block 1, it was a problem that one group of sentences contained more short words compared to others, but hopefully this will be balanced out across all blocks.

Gibson (2006) also warns against using bi/trigrams to estimate syntactic expectations, because an expectation may be initiated farther back in the input string.

Another theory is the end effect. There may be an effect at the end of the sentence, where the predictability of the baseline sentence should be high, so the reading pace will go down, but for GP and CS, it will break the predictability, so the reading pace will go up. Therefore, the last words of the sentences have been tagged 3, 2, 1 to see if there is an effect here.

4.5.6 Stimuli

For a complete list of stimuli used in experiment 2, see appendix A and B. Languages used are Norwegian Bokmål and English (no words that indicate that it is American or British English). The Norwegian script variant of Bokmål uses the Roman script, like English, but extends it with the three letters æ, ø and å. These three letters have been avoided in the stimuli.

Target match is never in the same position in search set and target set.

I avoided cognates in the memory test. Cognates share meaning and orthographic forms, phonological forms, or both in the two languages (Jared, 2015). Numerous studies have shown a quicker response to cognates than single-language words, see (Jared, 2015) for an overview.

Memory test B has a translation of a cognate, frog - frosk, and also one translation cognate - kindergarten. For more in-depth comments on stimuli, see chapter 3, and section 4.4.2 in chapter 6. Point after CS includes 3-5 words. GP sentences used in experiment 2 were created using the definition worked out in chapter 2.

I also took into consideration the results of the pilot study, where there were some outliers of the sentences within the category of GP. Also, as mentioned in the acknowledgements, professor Helge Dyvik was a great help in the finalizing of the GP sentences.

The sentences used for experiment 2 have 7 or 8 words per sentence. Due to an error in SuperLab, one sentence (sentence 10 in block 3) has only 6 words. In block 1, sentences 6, 8 and 11 have eight words. In block 2, sentences 4, 9, 11 and 12 have eight words. In block 3, sentences 5, 6, 8 and 9 have eight words. In block 4, sentences 4, 5, 7, 8 and 9 have eight words. All others have seven words per sentence. As we can see, sentences with eight

words varies across blocks, and makes them less comparable. But forcing all sentences to have seven words (or all to have eight) would involve serious compromise regarding GP, or make CS impossible if the sentence were to be grammatical in English as well, see more details on compromise on materials in chapter 3.

4.6 Data Validity

The quality of an experiment relates whether the measures are generalizable, reliable, and valid (Lanza, 2008; Eckert, 2013). This is ensured by choosing a method that best tests the factors to be tested in a sound and robust way.

Within the field of psycholinguistic studies, one has to be aware of the relationship between the test situation an a real-life situation. *Observers paradox* is the participants awareness that he or she is being studied, which may have an effect on their linguistic behavior, and thus make the data less representative of real world reactions (Wray and Bloomer, 2012). In the same way, ecological validity refers to the relationship between the study and the real world, it refers to the artificial test setting and the real world.

The process of reading in a word-by-word experiment is different from the process of everyday reading the newspaper or a book. On the other hand, it is impossible to study everyday reading without observers paradox or questionable ecological validity. Head-mounted eye-tracking methods are bad in this sense, as they are mounted onto the head, not allowing the participant to move his or her head freely.

Valid data is data that measures what it is supposed to measure, in this study the process of reading different types of sentences. Now, if the process of reading in the test situation is too far from the natural non-tested process, the data will not be valid.

Studying CS in a reading study is more unusual than studying it in spontaneous speech. It is constructed in an unnatural way that may not occur in natural reading or writing. Still, it has the value of testing the bilingual lexical access in bottom-up information retrieval. As stated by (Bullock and Toribio, 2009), it is a methodological problem within the field to study CS without compromising the phenomenon through replicating, inducing, or manipulating it. I have chosen a different approach to studying CS, with some elements from language switching. See definition in 1.3.

The validity of a study refers to the correct application of research conventions (i.e. methods, approaches, and techniques) to address the questions or issues you set out to study. For research to be valid, it must observe, identify, and measure what it claims to. Validation is a process of gathering evidence and constructing theoretical arguments to support the results and their interpretation(Lanza, 2008).

4.7 Method for analysis

To assess the probability that differences in data (e.g. differences in RT) are driven by differences in the values of their independent variables, (e.g. GP or CS), rather than just by chance, statistics on gathered data is used (Moyer, 2008). Gries (2013) gives two main reasons for applying inferential statistics: to arrive at better estimates of population parameters, and to test hypotheses and separate random/accidental from systematic/meaningful variation. First, let's review what we are testing. My main hypotheses aims to test whether the combination of GP and CS is significantly faster than predicted by an additive effect or not, and whether there is an effect of L2 and set length in working memory.

For the main analyses of the data that will be gathered, I will use linear-mixed models. But descriptive statistics can also be good indicators of results. Descriptive statistics do not dig as deeply as a linear-mixed model does, but they can be clear indicators of trends in the data. Examples are mean, median, standard deviation from mean, or range. I have studied some structures of GP, hoping to draw conclusions on the effect of GP in general, as based on the definition.

Before I present what the model does, I will present the different parts that make up the model. The model uses fixed and random variables to estimate the effect of a fixed variable (e.g. RT) on independent variables (e.g. sentence type). A variable is anything that has a value that can vary, such as the number of years of L2 study or the participants native language (Baayen, 2013). Variables can be continuous or categorical. Independent variables are the factors that vary, such as sentence type (BL, GP, CS or GP+CS). Dependent, or fixed, variables are what is measured in the experiment, such as RT or correct responses. In experiments, it is researched how the dependent variable affects the independent variable. Let me give an example from a model I will use on the SPR-test. For each of the participants, there will be 12 test sentences, with 7 or 8 words per sentence. These sentences will be different types of sentences, with variance within the group, and the words will differ in length and category. Both the participants and the sentences should ideally be sampled randomly. Sentences and words are random effect factors, which will be treated as sources of random variation in the data. A slope will be calculated, with an intercept. Participants will read the sentences at different paces, and fast participants may on average read a GP sentence faster than a slower participant reads a BL. This does not mean that there is no effect of GP, it means that both participant may show an effect of GP, but that the intercept for each participant is different.

Random factors that I have taken into measure in the analysis are participants and items (word or sentence). I measure the effect of the variables, so there is both a fixed variable showing a fixed effect, and a random variable giving a random effect. Using both random and fixed effects controls the risk of underestimating variance, which is a problem for fixed-effects models (Baayen, 2009).

I will use ”R” and its ”lmerTest”-extension package to carry out the analysis. The recent development of this package allows for estimation of degrees of freedom through Satterthwaite approximation, and also allows for esitmataion of the size of effect and the structure of effects through linear regression models. It will robustly estimate significant of factors in the model, controlled by using an ANOVA on the model.

Linear-mixed models are ideal because they take into consideration both variation in items (words or sentences) and participants (Baayen and Milin, 2010). Using linear mixed models allow for variance from other factors that may not be related to the experiment. Certain words can be faster, certain participants can be faster. Reasons to why some words are quicker to read or more recognizable are not always known. We know that concrete nouns often are easier to recognize than non-concrete nouns. Many speculate that frequency, predictability and familiarity play a role. We can control word processing to some extent, for example by using concrete nouns only as stimuli in the memory test, as they have been proven to be easier to process (Gathercole, 2009).

In R, random-effect factors are specified between parentheses. The notation (1 | Sentence) indicates that the model includes a random intercept for Sentence, where each sentence has an expected starting point or average (Baayen and Milin, 2010). This allows for the possibility that some sentences might be more difficult or more complex, leading to longer reaction times across all words in the sentencce and across all participants (Baayen and Milin, 2010). The notation (GP*CS|Participant) specifies the random effects of GP and CS for participants, making the model take into account that participants may not react identically to GP and CS or their interaction, as indicated by ”*” in R. This can also be used over sentence, as the effect of GP and CS is not identical in each sentence. The dependent variable shows differences that are related to fixed variables. Sources of extra variance are typically participants and items (word or sentence in my case). Other variables may also affect the outcome of the analysis. Other non-controllable factors are the weather, time of day (of course controllable to some extent, data was only gathered on weekdays between 8-17). Gender has not been used as a factor in this experiment. Also, we can not control how much the participant has slept the night before, and since I have been testing students, it is not unlikely that some participants may have been hungover. Random factor of participant should balance this out.

The quality of the models can be assessed. A qq-plot can show the goodness of fit of the model, i.e. how much of the effects in the models that are explained by the model. Other assessments are degrees of freedom, it should correlate to the amount of factors in the model, that indicate that the planned design has been useful for explaining variance. This makes it harder to detect significance, but it lowers the risk of type I errors, and makes it more likely that results will replicate.

Each model will be evaluated with an ANOVA. Analysis of Variance (ANOVA) is a robust way of investigating whether two scores are as different as the hypothesis predicts (Wray and Bloomer, 2012). ANOVAs are ideal in exploring interactions between variables, as they

allow for both fixed and random variables. A type III is used with a sattertwait approximation to test the model for significance. Type III ANOVAs prioritize interaction effects in the calculation.

Chi-square tests for independence are used on two categorical variables, for example correct responses for condition of memory test. The chi-square will test whether the observed frequencies (for example correct responses) vary in the different versions of the memory test.

For the correlation, a Spearman's rank coefficient correlation test has been used (R: cor.test). It is a non-parametric measure of statistical dependence between two variables. It tests whether two ranks are similar (positive correlation) or move in opposite directions (negative correlation) (Wray and Bloomer, 2012) Comparing the language in memory test and the different conditions of the sentences in the SPR-test is tricky, as they are different tasks and yield different RTs. A spearman's rho will indicate a correlation between for example an increase in RT.

Chapter 5

Data and Results

This chapter will present the results from the experiments. First, I will briefly go over the results from the pilot study, focusing on the comprehension task that was discontinued in the newer versions. Then we will take a brief look at results of Experiment 1. Note that the results from this experiment have not been analyzed in depth, but are used as an indication of task compliance. Then I will present the results of the different analyses of Experiment 2, present the results of the SPR study, then the memory test, and finally look for a possible correlation between the SPR-test and the memory test.

5.1 Results of pilot study

5.1.1 Comprehension Task

First of all I will present the results of the test that was discontinued (see 4.3.3 for discussion). This task was to pick the version of a sentence that corresponded to the sentence they had just read, with four topicalized versions of the sentence to choose from.

Here we tested for a correlation between GP sentences and amounts of errors made. Pearson's chi-square tests with Yates continuity correction was used to estimate significance. There is a significant association between GP sentences and amount of errors made in comprehension questions:

Chi squared: (df= 1;N= 488) = 22.12,p <0.001***
Effect size small/medium (Phi <0.22)

However for CS, there was no significant association between code switching and amount of errors:

Chi squared: (df= 1;N= 488) = 1:69,p <0.19
tiny effect size(Phi <0.06)

Table 5.1: Pilot Study:Correct and Wrong Answers by Garden Path and Code Switching
No GP, No CS:
Correct: 100 (82.0%)
Error: 22 (18.0%)

No GP, CS:
Correct: 98 (80.3%)
Error: 24 (19.7%)

GP, No CS:
Correct: 81 (66.4%)
Error: 41 (33.6%)

GP, CS
Correct: 69 (56.6%)
Error: 53 (43.4%)

As can be seen in the table 5.1, the main effect is by GP, but CS does play a marginally larger role when in the GP.

The results shown in table 5.1 show that the effect of GP and CS together is not additive for the comprehension questions. CS does not affect correct responses in comprehension questions, but GP does. In all answers, there is correct responses higher than chance. Chance would here be 1/4 = 25% correct. This indicates that task compliance is assured - the participants performed much above chance.

The results for comprehension includes subjects who had problems reading the experimental sentences on time. In this experiment, the comprehension questions were presented directly after the sentence. This may have affected the subjects to read very carefully, and maybe use memory strategies.

5.1.2 Results of Self-paced Reading

As stated before, for this part of the test only 36 out of 62 participants performed to criteria. As discussed in section 4.6, the combination of a very challenging comprehension task and SPR proved problematic. Some participants were excluded because they did not show that they had read and understood the sentences in the comprehension task. The participants who were not excluded solved the task in over 400 ms, but less than 3000 ms per half.

The reaction times were analyzed using a linear mixed model with random effects for subject and sentence. The regression is:

RT = 1396.2 + 177.3 (GP) + 153.3 (CS) - 135.7 (GP+CS)

An ANOVA shows that GP adds a significant effect:

$F(1, 445.2) = 6.32, p<0.012$ *

The effect of CS is approaching significance:

$F(1, 444.8) = 3.86, p<0.05$

The interaction effect shows that when both GP and CS is combined, the effect of CS is canceled out:

$F(1, 445.7) = 2.43, p<0.120$-

There is no interaction, and the effect is not additive.

There is a risk that fast readers compared to slow readers in this experiment represents different populations of bilinguals. It could be that the fast readers were also better at comprehending. This suggests that the status for bilingualism should be tested more explicitly than was done. However, the obtained results indicate that parsing L1 and L2 for faster readers may rely on the same parsing mechanism, rather than separate grammars, since CS did not slow down reading times when a reparse was necessary in the GP sentences.

5.2 Results from Experiment 1

Although the data was not balanced, there are still some conclusions to be drawn that are helpful indicators for the third and final experiment.

Participants completed the task as expected. I gained more experience in conducting experiments, and I conducted the experiment with all participants myself, to better control for variance in instructions. In the figure 5.1, we see that there is a difference in RT between

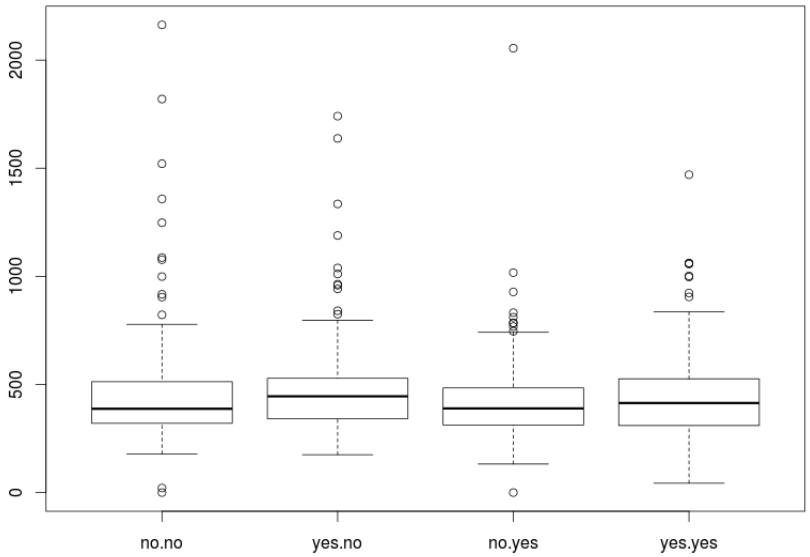

Figure 5.1: Experiment 1: Box plot of Reaction Time of GP*CS
First box is baseline, second is GP, third is CS, fourth is GP and CS

the baseline and GP sentences, but that CS-sentences seem almost as quick as the baseline. There is a slight increase in GP+CS.

This box plot 5.1 shows the overall general effect, and as explained in section 4.4.2, the sentences are not controlled for length. This part of the experiment does not contain unbalanced or otherwise unvalid stimuli, but since the memory test was unbalanced, it is not possible to do an analysis of correlation between memory span and reading pace on a per participant basis.

The RTs presented in figure 5.1 also show that the participants solve the task as expected, with clear outliers that can easily be spotted and excluded. For this experiments with 48 participants wth 8 sentences with on average 8 words per sentence adds up to approximately 3000 data points. Out of the approximately 3000 data points, there are 10 outliers with a RT over 1200 ms visible to the naked eye.

5.3 Experiment 2

This is the main part of the analysis. The pilot study had problems with task compliance, Experiment 1 had problems with unbalanced stimuli but showed great improvement in task compliance, and the final Experiment 2 has both valid data from balanced stimuli and high task compliance.

The experiment consists of two parts: first a non-stationary, moving-window, word-by-word self-paced reading experiment, then a memory test of differential working memory in

L1 vs L2.

5.3.1 Participant Demographic

All together, 74 participants performed one of the memory versions, and one of the blocks of SPR. The participants were between the ages of 19 and 31 and had Norwegian as their first language. There were 24 people taking memory test A and 24 taking memory test B, and 26 taking memory test C. There were 19 participants performing block 1, 19 participants performing block 2, 18 participants performing block 3 and 18 participants performing block 4. Out of the 74 participants, 67 performed to task, where 5 were excluded due to reading/writing disorders, one did not complete the experiment, and another one reported after the experiment that he/she did indeed not have Norwegian as their first language, even though he/she had been asked before the experiment. So the distribution of the participants who performed to task is as follows:

Memory test:
A: 22
B: 23
C: 22

SPR:
Block 1: 17
Block 2: 16
Block 3: 17
Block 4: 17

5.3.2 Data

Memory tests A and C consist of 528 data points, and B of 552 (one more participant). Each of the blocks in the SPR test contain from 1478-1531 data points, one block has 16 participants, the others have 17. Block 4 consists of more words. The control questions with the two immediate sentences add up to 1037 data points, same across all blocks.

A t-test and standard deviation calculated in R is used to select the 99th percentile, so reaction times under 200 and over 1201 are excluded. In block 1, this means that 48 data points are excluded in the lower bound, and 27 of the upper bound. In block 2, 28 are excluded in the lower bound, and 18 of the upper bound. In block 3 it is worth noting that many participants were performing near to 200 ms, with 20 responses in the range 190-200, but they were excluded, so 56 are excluded from the lower bound, and 33 of the upper bound. In block 4, 35 are excluded in the lower bound, and 28 of the upper bound.

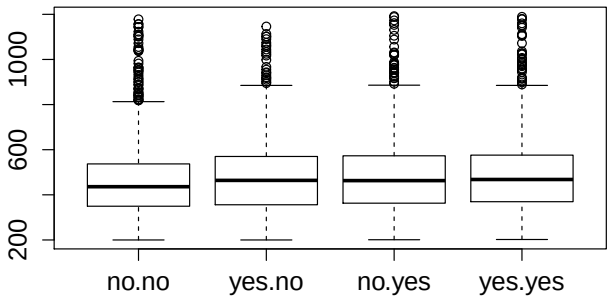

Figure 5.2: Experiment 2: Box plot of RT of SPR for all blocks
First box plot is baseline, second is GP, third is CS, fourth is GP+CS

Memory files were tagged manually in excel for set length, language of the match word, and translation direction (for B and C). RTs under 350 were removed, excluding 48 data points from version A, 101 data points from version B (of which 100 were no responses), and 17 from version C. The span of the remaining reaction times now span from 668-2940 in version A, 746-2971 in version B, and 621-2999 in version C. Something that catches the eye is that in version C, all of the top 20 quickest responses are yes-responses, as predicted by (Sternberg, 1966), correct yes-responses are faster than correct no-responses. Upper time-out was set to 3000 ms in SuperLab5.

Sentences were then formatted so that in sentences with 7 words, the first word was removed, and in sentences with 8 words, the first two words were removed, so that all sentences have 6 data points (given that none have been excluded in previous file handling rounds). After removing the first, or the first and second word of the sentence, the complete data set now contains 4623 lines of data, of which the distribution of GP and CS is as follows:

GP: CS:
no: 2336 2308
yes: 2287 2315

This shows a fairly even distribution of data points between the categories.

5.3.3 Self-Paced Reading Experiment

After all the formatting is done, the data sets have been cleared for analysis, some files have been tagged, and we have decided on which models to use, statistical analysis can begin. The first indicator to look at is the reaction times for the SPR-test. Figure 5.2 is a box plot of the RT for the four conditions, for all blocks combined.

As we see in figure 5.2 there is a slight difference between the baseline (first box) and the

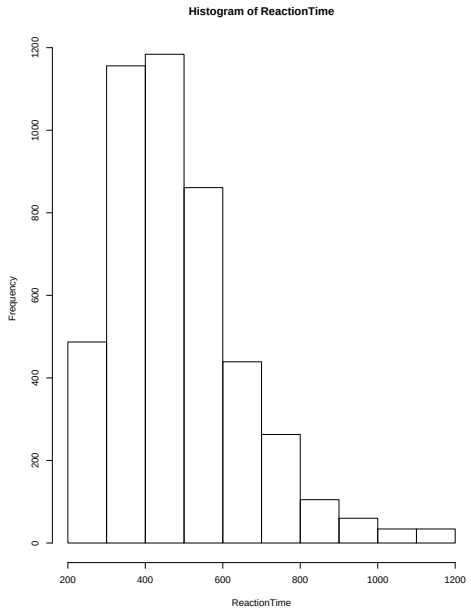

Figure 5.3: Experiment 2: Histogram of Reaction Time in SPR-test
Histogram showing the distribution of reaction times for the SPR-test across blocks, after outliers have been removed.

rest that is visible to the eye. We now have an indication of the results to come. But before we can conclude on any firm results, we must do a deeper analysis of results.

The histogram 5.3 shows that reaction times are left skewed and centered around 400 ms. This corresponds to the mean reaction time of 483 ms. As we see, there are not many reaction times over 1000 ms, so we could maybe have been more restrictive in excluding upper-bound reaction times. Outliers were selected as more than 2 standard deviations from the mean.

As we saw in section 4.5.5, I formatted and tagged the files of the SPR-test. One file had removed either first or first and second to make all sentences 6 words and thus more comparable. From this, we get a general idea of effects. The analysis of this file can show us the effect in the sentence as a whole, where word length may be better balanced than in the file tagged for point of GP/CS or end effect of the sentence where only three or four words are analyzed, and thus an imbalance in word length may affect the outcome of RTs.

5.3.4 Analysis of difference in medians

Effect of medians:
lmer: ReactionTime - GP * CS + ((GP * CS) | Participant) + ((GP * CS) | Sentence)
Number of obs: 802, groups: 67 Participants; 12 Sentences

$$RT = 459.2 + 16.1(GP) + 15.2(CS) - 19.4(GP + CS)$$

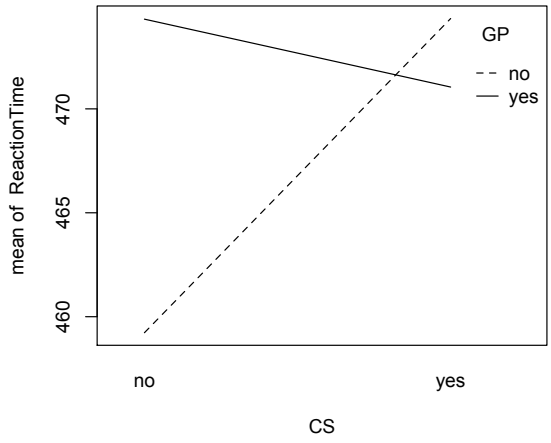

Figure 5.4: Experiment 2: Interaction effects of model of medians

ANOVA: GP:CS $F(1,180.2)= 4.01$, $p<0.04$ *

A Linear mixed-model of median of reaction time to GP and CS, with random effects GP and CS over participant and sentence. This allows the participant to react differently on GP and CS, and also there might be a difference in effect of GP and CS in the different sentences. This is done on the set of 6 words per sentence-file. Results show that CS does not have a negative effect on GP reaction times.

Figure 5.4 shows the effect of CS, as well as the effect of GP, and then how the two interact, showing the interactive effect is not additive, and that GP is actually a tiny bit faster with CS.

5.3.5 Point of GP and CS

We tagged the files for point of GP and CS, and point of baseline (where the GP or CS would have occured, or occurs in corresponding versions of other blocks). This means that the two words before the effect is expected are tagged -2 and -1, and the two words after the point at which the change happens (or doesn't happen) are tagged 1 and 2. This is the point at which one expects to see a increase in RT.

A linear mixed model of the extracted two words after the CS or GP occurs (or doesn't occur) we get these results:

ReactionTime - GP * CS + ((GP * CS) | Participant) + ((GP * CS) | Sentence) + (1 | Block)

RT = 459.2 + 25.7(GP)** + 35.3(CS)** - 40.9(GP + CS)*

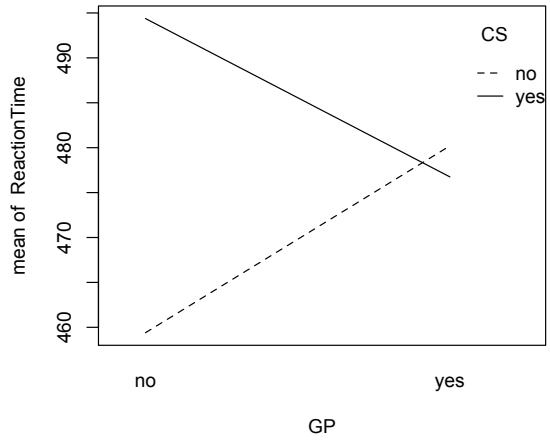

Figure 5.5: Experiment 2: Interaction effects at Point of GP/CS

ANOVA of interaction: $F(1,11.3)=6.2$; $p<0.05$ *

This analysis uses GP and CS as random effects on participant, meaning that participant may react individually to GP and CS, and the same for each sentence. Also, there might be a difference between blocks. There is a large span in what a GP sentence can be, and also we have seen a difference in RTs for sentences within the group of GP in the pilot study.

GP adds a significant extra reading time of 25.7 ms at the point of GP, CS adds a significant extra reading time of 35.3 ms at the point of CS, and GP and CS combined subtracts 40.9 ms to expected reading time, also significant.

The figure 5.5 shows the effect of GP (note that this graph has GP as a contrast, as opposed to figure 5.4 that has CS as contrast). The graph shows the effect of no GP, no CS, CS only, and the interaction effect.

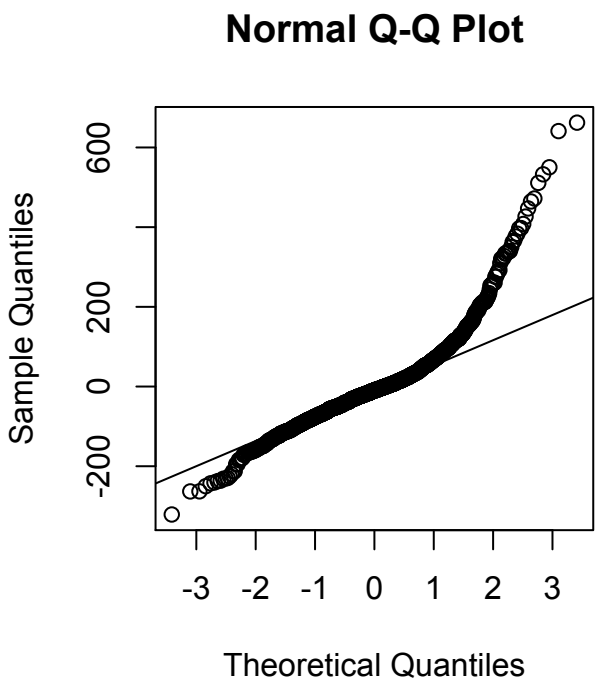

Figure 5.6: Experiment 2: Quality of model - qplot

The figure 5.6 shows the theoretical quantities if they were to follow the normal distribution. The goodness of fit shows a well fitted model up to the second quartile. From -200 through 200 it is near perfect.

We can also look at position 1 (after point of GP/CS/baseline) in isolation.

RT = 460.2 + 16.7(GP) + 33.8(CS) (**) - 37.4(GP + CS) (**)

In this analysis, there is no significant effect of GP. This may be because the effect comes later on in the process, but for CS, the effect is more immediate. This is in line with the first analysis, where there is a higher effect of CS (35.3) than GP (25.7). Also, it may be because many one or two letter words occur in GP sentences at this position.

For comparison, let's look at position 2 after point of GP/CS/baseline:

RT = 458 + 34.9(GP) (**) + 36.7(CS) (**) - 44.7(GP + CS) (**)

Here we see that the effect of GP is clearer in ms.

Adding pos 1 and pos 2 allows us to get a better controlled model, which lowers the risk of type I errors (i.e. reporting significance that has been inflated by an unmotivated df).

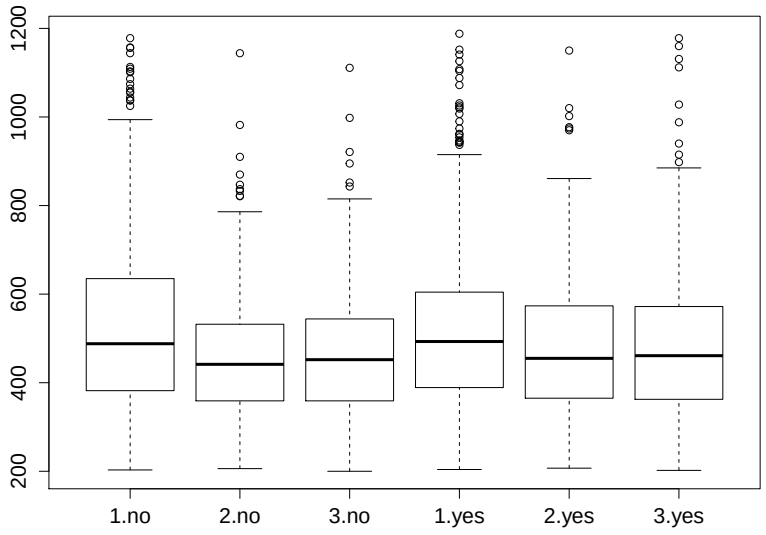

Figure 5.7: Experiment 2: Box plot of RT of end positions for CS
first three (no) is no CS, positions 3, 2, 1 are counting down to sentence end. 1 is last word of sentence

5.3.6 End Effects

We saw that in the analysis of point of GP and CS, there was a more immediate effect of CS. Maybe when analyzing the end effect, the opposite will be true? CS is an immediate effect, but GP shows effect also at the end of the sentence.

We tagged the files for the end effect, tagging the three final words of the sentence, counting down 3, 2, 1. For this, in the baseline we expect to see no increase, or even a decrease in RT for the baseline, as the sentence should become more predictable as one reads, and one should then be able to keep a stable pace, or even increase it as one learns the structure of the sentence.

The GP sentences should then show effect by increase of RT at sentence end due to unpredictable structure.

An linear regression model of reaction time by GP and CS, with random variables GP and CS effect on participant, and sentence, has been done. Variance across blocks, according to the random effects analysis, is 0, so block is not included as a random variable.

Formula: ReactionTime - GP * CS + (1 | Participant) + ((GP + CS) | Sentence)

RT = 482.1 + 17.8(GP)* + 19.9(CS)* - 15.4(GP + CS)

ANOVA: GP $F(1,37.42)=3.69$, $p<0.06$

The effect of GP is near significant. No significance is found for CS or interaction. The results can be visualized through box plots.

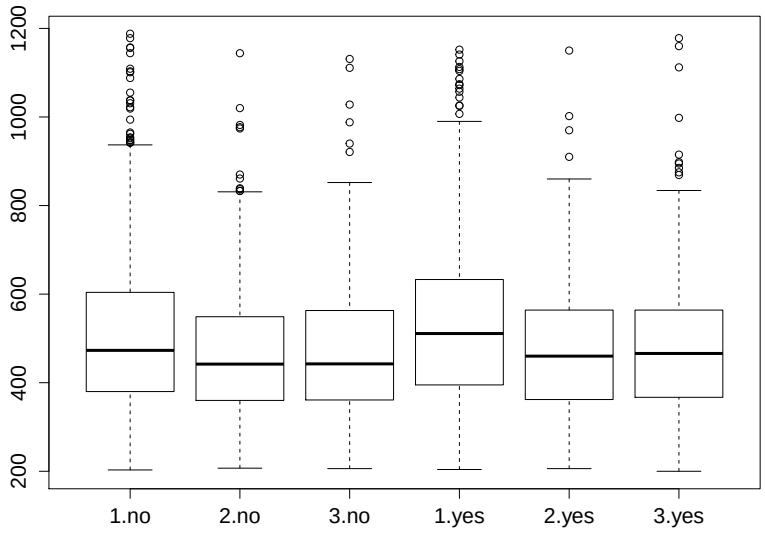

Figure 5.8: Experiment 2: Box plot of RT of end positions for GP
first three (no) is no GP, positions 3, 2, 1 are counting down to sentence end. 1 is last word of sentence

Comparing 5.7 and 5.8, we see a more stable result for end of CS, with slightly higher RT for all three words of CS. The last word is the slowest of the three for both box plots. This may be explained through the reading process of a sentence. From the non-stationary moving-window self-paced reading, the participant can tell that this is the last word of the sentence, and may take a little extra time on it. It can also be that the participants wrap up the sentence when it is finished and the full understanding of the sentence includes the integration of the last word.

5.3.7 Comprehension Control

For Experiment 2, four comprehension questions were used to ensure that participants were in fact taking in information while reading the sentences and not just clicking the button. The data show some outliers, as participants may have clicked further expecting only one word, not a question. There was no pre-warning indicating a question was coming up in the experiment, but the participants were informed orally before the experiment started that questions followed only some of the sentences, and as they are randomized, they might feel a bit sudden. This can be better controlled for in upcoming versions by for example marking the arrival of a question after the sentence is read, so that the sentence will still be read by the participant not knowing whether a question will follow it, but then after the sentence is read, the participant will be notified that a question will now follow.

Results are not to be analyzed in depth, but to be used to show that participants did indeed read sentences, and not just press buttons at random.

Table 5.2: Memory Test: mean RT and mean RT for correct responses for each version
A - mean RT: 1618 ms, correct: 1582 ms
B - mean RT: 1711 ms, correct: 1691 ms
C - mean RT: 1688 ms, correct: 1634 ms

There are four questions per participant. Per block, this adds up to 68 data points, all together 272 data points for all four blocks. Answers should be yes or no, but some have pressed the continue button, accidentally skipping the question. This was at the most 8 continue-presses in block 1, and the least 3 continue-presses in block 4.

Out of the 272 data points, answers are distributed followingly:
22 continue, 8%
220 correct, 81%
30 error, 11%

As we see correct responses are way over chance, which would be 50%, or 136 correct, even when considering missed questions as errors. This means that we can safely assume that participants did indeed take in information as they were performing the SPR-test.

5.3.8 Memory Test

The memory test, as we recall, comes in 3 conditions. It has not been through as many tagging processes in the file formatting process as the SPR-test, and we can see some indicative results.

First of all, we see that the differences between the files are not very large. As expected, for the distractor version (B) there are more errors, and thus fewer data points to analyze.

Correct responses:
A set: 395 correct (85%), 72 error (15%) total: 467
B set: 233 correct (52%), 218 error (48%) total: 451
C set: 429 correct (84%), 82 (16%) total: 511

As we see, there are no large differences in errors between sets A and C, but there are slightly more correct responses in C-set. Percentagewise, the different is not very big, as there are more responses all together in C-set. More RTs have been excluded from A- and B-set than the C-set. The B-set has many more errors than the other two. This is expected, as the translations can be interpreted as a distractor effect.

RTs were recorded. Mean RTs for memory versions A, B and C are presented in the table 5.2 If we extract RT for only correct answers, mean RT across blocks was 1633, with

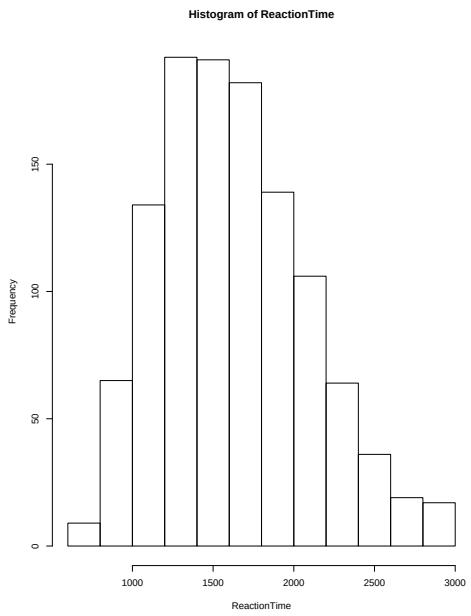

Figure 5.9: Memory Test: Histogram of Reaction Time for correct responses
The histogram shows correct responses for all versions of the memory test

x	A:	B:	C:	D:
A:	46	54	42	44
B:	41	40	39	36
C:	48	51	49	44
Sum:	135	145	130	124

Table 5.3: Memory Test: Correct responses per position A B C D

correct response times for each version presented in the table 5.2. As we see, version A (no translations) is slightly faster than the other two, and also faster than the mean of all.

The histogram 5.9 shows the distribution of reaction times. Reaction times are clearly centered around 1500 ms, which corresponds to the mean RTs of thee different blocks. The distribution is near to normal distribution.

We controlled for each position of the match word in in the search set. The search always has 4 words, and the match word might occur in each of the positions, A, B, C, or D. There is a difference in correct responses between these four positions, as shown by table5.3.

In the table 5.3, using Pearson's chi square test, there is no significant effect of position of the match word (sum of all):

Chi squared = 1.7753, df = 3, $p<0.62$

However, for all three versions, most correct answers are found in position B. There is an effect of length 8 in English across all versions of the memory test. L2 yields a higher RT for decisions in all versions. The effect of length 8 is significantly slower for L2 than L1 across all three versions of the memory set, as shown in table 5.4.

Table 5.4: Memory Test: Set length related to match language
A: df = 2, N = 395, p = 0.01463 (*)
B: df = 2, N = 233, p = 0.04889 (*)
C: df = 2, N = 429, p = 0.012664 (*)

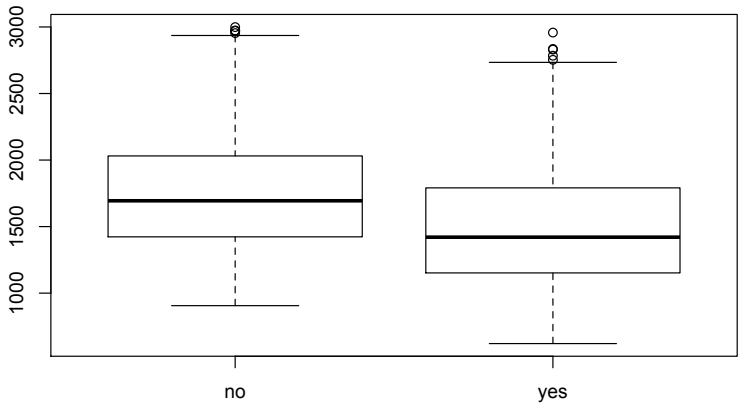

Figure 5.10: Memory Test: RT of Correct Responses Yes or No

There is also a difference in RT for the distribution of correct responses yes or no.

There is a quicker response time for correct yes responses presented in figure 5.10, which is in line with Sternberg's (1966) prediction. This may be because when the participant reads the through the search set looking for a match word, in a yes-answer the participant can click as soon as he or she sees it, but in the no-answer, the participant will have to read the whole set before answering no.

I have looked at amount of correct responses across sets, and I have also looked at the effect of position of match word (ABCD) in the search set, but is there a different in correct responses per set length?

We see that there are more correct responses for the 4-set as one may expect. But the difference between sets is not enormous, as shown by table 5.5.

As we see in table 5.5, the 4-set has the most correct responses, with 28 more correct responses than 6-set, which has 14 more correct responses than the 8-set. This shows that there is an increase in difficulty for set length.

The box plot 5.11 shows reaction times for each of the set lengths for each of the lan-

Table 5.5: Memory Test: correct responses per set length
4-set: 408 correct
6-set: 380 correct
8-set: 366 correct
Sum: 1154

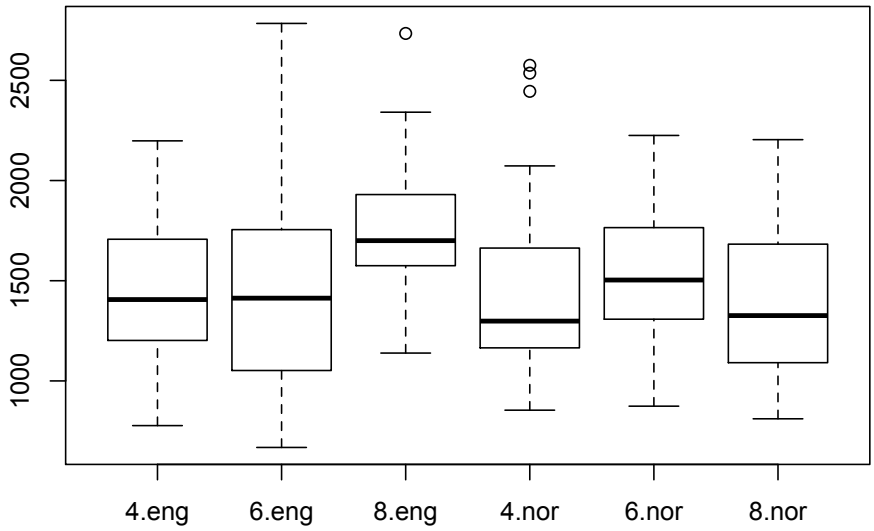

Figure 5.11: Memory Test: Box plot of length of memory set and language for A-set

Table 5.6: Memory Test: correct answers by language and set length
4 eng: 88 correct, 6 eng: 116 correct, 8 eng: 67 correct
4 nor: 97 correct, 6 nor: 63 correct, 8 nor: 104 correct
English correct: 271
Norwegian correct: 264
Sum of yes-corrects: 535

guages. It is comparable to the table 5.5 in that we see an increased effect for 8-length. In table 5.5 we saw that there are fewer correct responses from set length 8, but the box plot 5.11 reveals that the effect of the set length is only increased (in reaction time) for set length 8 and language English.

The table 5.6 presents correct yes-responses by language. For set length 4, there are slightly more correct answers for Norwegian. For set length 6, there are almost twice as many correct responses for English. For set length 8, there are many more correct answers for Norwegian. All in all, there is no big difference in correct answers per language.

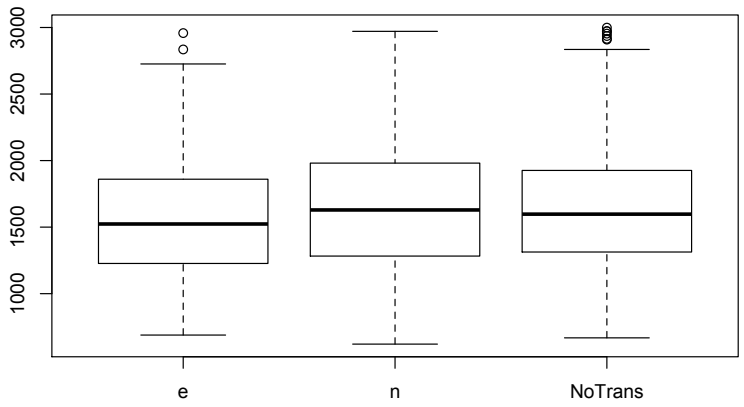

Figure 5.12: Memory Test: Box plot of RT for translation direction
The first box -e shows RT for translation direction from Norwegian to English, second box shows from English to Norwegian, and the third shows RT for no translation.

5.3.9 Balancedness

The two languages in the memory test can also be an indicator of how balanced bilingual they are.

I have tagged the translation direction of the translated words, whether it is translated from Norwegian into English (-e) or from English into Norwegian (-n). There is a difference in reaction time between translation direction.

The box plot 5.12 shows a faster RT for translations with a Norwegian match word translated into English. Translations from English into Norwegian does not give a quicker RT. Translations only quicken response time when they are translated from Norwegian into English.

If match language and participant were statistically independent of each other, they would all follow the line. But we see that while some participant show near or no difference in match language, some participants do. It is quite clear to see how the participants react, this can be used as an indicator of balancedness. Also, participants that are unbalanced can be selected based on these plots, to see how it is in correlation with the SPR-study.

In figure 5.13 Association plots shows how association per language, with *black* indicating more associated than expected, *red* less associated than expected for each participant.

Balanced participants are participants who do not deviate from the line. Examples in 5.13 are participants 15, 17 and 25. Participants who deviate from the expected frequency are considered less balanced. It is a relative measure of expected frequencies. Participants 15, 17 and 25 are exactly as expected in this data set. 14 and 39 are extreme in that they are under represented for Norwegian and over represented for English. 26 and 41 are extreme in the opposite direction.

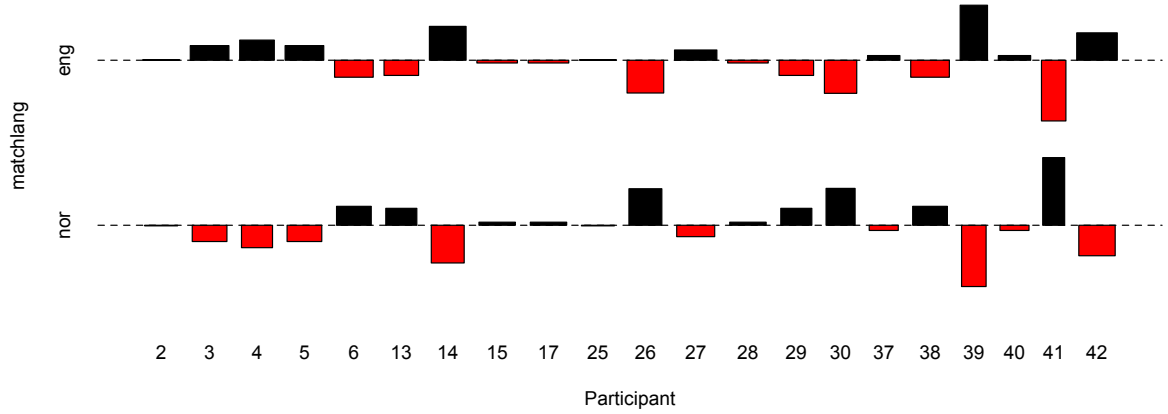

Figure 5.13: Memory Test: Association plot of A-set

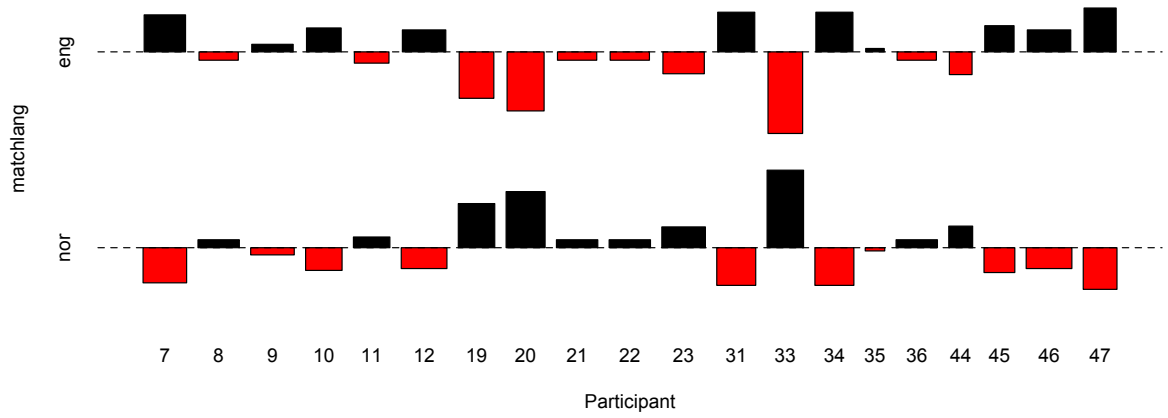

Figure 5.14: Memory Test: Association plot of B-set

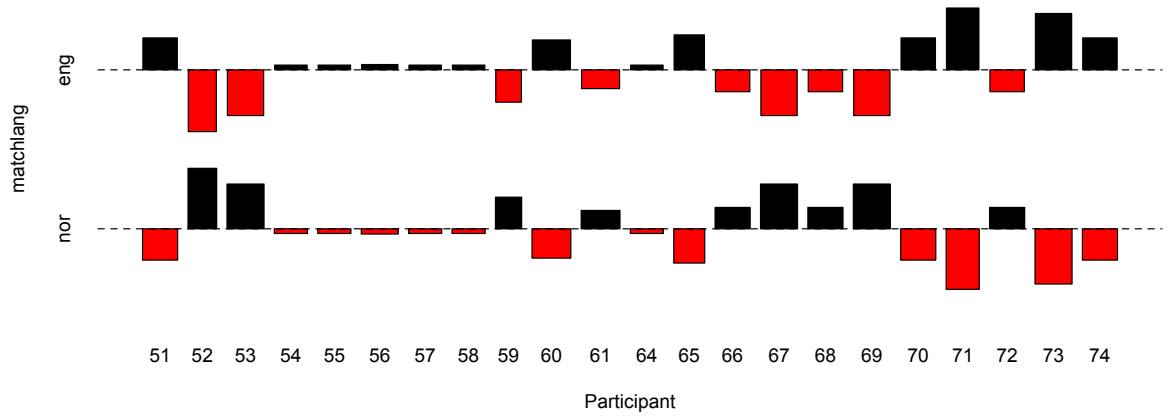

Figure 5.15: Memory Test: Association plot of C-set

Because the B-set has less data points than the other two, there is a greater risk of unbalance here. The C-set has more participants with a slight unbalance.

5.4 Correlation between Memory Test and SPR

There is a risk that fast readers compared to slow readers in this experiment represents different populations of bilinguals. It could be that the fast readers were also better comprehenders.

One indicator of per-participant results is the individual reaction times for the two languages. Here, the match language is used as a factor. Match language refers to the language of the match word. It the word that they will be presented with in the memory set, and then find in the search set. Using the match language as a factor can tell us whether it is easier for the participants to recall a word from Norwegian compared to English. Box plots on a per-participant basis have been added to the appendix. They differ from the association plots presented in the previous section in that they show differences reaction times rather than correct responses.

5.4.1 Spearman's rank correlation rho for SPR and Memory Test

The Spearman's rank-order correlation is the non-parametric version of the Pearson product-moment correlation. Spearman's correlation coefficient measures the strength of association between two ranked variables (Wray and Bloomer, 2012).

Data from the SPR-test and the memory test were combined in an excel sheet, with per-participant information. The per-participant RT scores were assembles using R aggregate-function.

Factors were: Participant in both test. Memory test: match language English, Norwegian or no match. SPR: baseline, GP, CS and GP+CS.

The analysis will test for a correlation between match language and each of the four conditions of the SPR test. Mean RT for match language (and no match) were combined with mean RT for each of the four SPR conditions, baseline, GP, CS and GP+CS.

No correlations for match language combined with any of the four conditions were found.

Of particular interest is whether there is a correlation between match language English and CS. A significant correlation between response time on match language English and response time on CS sentences could tell us that bilingualism in working memory affects processing of CS. The test can tell whether the results of the memory test, quick or slow, correlate to quick or slow responses on CS sentences in the SPR test.

A Spearmans rho of match language English on CS yielded no significance.
rho = 0.1270756, p = 0.3282

No significance was found here either. There is no correlation between response times of English match words and CS sentences in the SPR test. No correlation was found between any of the 12 combinations of factors between memory test and SPR test.

5.5 Summary of results

The pilot study showed a significant increase in reading pace for GP sentences, and CS sentences, and the interaction effect cancelled out the effect of either GP or CS. The type of sentence had a significant effect on correct answers in comprehension questions.

Experiment 1 showed higher task compliance, and introduced a new memory test.

Experiment 2 replicated the results of the pilout study in that it showed a significant increase in reading pace for GP sentences and CS sentences, with no additive effect, where the effect of CS seems cancelled out. An analysis was carried out at the point of GP/CS compared to end position, where the main effects of both CS and GP are found at the point at which they occur, and not at the sentence end.

The memory test showed an increase in RT for set length 8 and English. There was varying task compliance between the three versions. Translations did not help the recognition process of match words, but slowed RTs slightly. There are more correct responses for category C - translations on match words.

No correlation was found in RT between memory test and SPR test. Match language in memory test was correlated to RT on the different types of SPR sentences (baseline, GP, CS, GP+CS) using a Spearman’s rank correlation coefficient, but no significant correlation was found.

Results support my H1 of the SPR-test: The combination of GP and CS is significantly faster than predicted by an additive effect.

Results of the memory test also supports my hypothesis H1 of language in working memory. I found there is a significant added processing load of longer memory sets (8) and L2 (English).

For the correlation effects, H0 was not discarded: H0: There is no correlation between performance on SPR-test and memory test. Correct responses and reaction times in memory test do not correlate with reaction times on SPR-test.

Chapter 6

Discussion

The initial aim of the experiments was to investigate the effect of CS on reading of GP sentences. GP and CS have been tested using two different methods of SPR, and two types of control measures. The pilot study laid the ground for combining the two linguistic phenomena in a new way, and experiment 1 and 2 further validated the results first found. The memory test has been introduced as a measure of bilingual balance in working memory.

The added processing load of GP has been proven in many studies (e.g. (Gibson, 2006; Sanz et al., 2013)), and I too have found an added processing load of GP. When combined with a CS, which also on its own added an extra processing load of change in bottom-up information flow, the effect was not additive. These results were consistent in both the pilot study and experiment 2 (data from experiment 1 has not been analyzed).

Theories presented in chapter 2 are in line with results presented in chapter 5. In the section 6.6, I will look at how Gibson's theory compares to the results I have found. The section 6.4.2 will be dedicated to discussing how the results of my memory test compares to the findings of Just and Carpenter, Caplan and Waters, and MacDonald and Christensen.

In this chapter I will delve into the results of the studies, and discuss what they may imply. I will discuss the method of using reaction time studies to investigate ambiguities, and how working memory is assessed. I will discuss the results of the pilot study compared to experiment 2. In experiment 2, I will compare the analyses of "endpos" to "point of GP/CS", to see where the effect lies, and also discuss the results of the different versions of the memory test. Finally, I will discuss what I have called the discount of bilingualism, a way of interpreting the results from the studies.

6.1 Discussion of methods

One confound of running an experiment testing an ambiguous sentence phenomena such as GP is that a participant might not discover the effect of GP. The benefit of a GP sentence, as opposed to merely ambiguous sentences, is that there is only one correct parse. However, the participant might read the words, but in its mind categorize the sentence as ungrammatical,

and thus miss the GP effect.

This is controlled for in different ways. Firstly, we saw in the comprehension test following experiment 2 that there is a high percent of correct responses, (81% if considering missed questions, or 88% if considering correct versus errors). Secondly, as I explained in section 4.7, in the lmer-model, slopes are calculated on a per-participant level, comparing the individual participant to the baseline, calculating for effects. This is a very robust way of statistic analysis, especially when combined with an ANOVA to control for intercepts. Finally, I have carried out analysis both at the point of GP/CS and at the sentence end. I found stronger effects at the point of CS/GP. If the participant had ”given up” on making sense of the sentence structure, he or she would have clicked speedily through the sentence. This would then have been detected when analyzing the last three words of the sentence, but no such effects were found.

Using reaction times to study ambiguities may involve a confound. Sentences like ”The horse raced past the barn fell” are so difficult that we sometimes need to be shown the correct parsing (Pinker, 2014). Reaction-time studies do not indicate which processes in the brain that are involved. Therefore, it might debatable whether an increase in RT in this study may be due to a confusion effect, and not a reparse. It is hard to find the ultimate cause of the observed differences, but when all data is accounted for, it excludes many possibilities.

I cannot guarantee that the sentence has been understood correctly, and this could to some extent be true of all sentences. The RTs can be caused by reparses, but the evidence I have found is only that some problem has been detected, and that problem stands in contrast to a baseline. The baseline and the conditions are thoroughly controlled.

As we have seen, there was an overall effect of GP, showing that most participants did react to it. An analysis at the sentence level shows responses within the group of GP, as well as an analysis on participant level may show whether some participants did not show an increase of GP. This variation is controlled for by including random factors for participant and sentence types in the analysis.

6.1.1 Assessing working memory

There is an ongoing discussion of whether the working memory resources employed during syntactic processing are different from the working memory resources used for other, more conscious verbal tasks (Gathercole, 2009).

In the pilot study, another way of assessing whether the participants had actually read the sentences was introduced. It proved to be a very challenging task, and did not test for bilingualism. The memory test introduced for experiment 1 and 2 used words as stimuli instead of sentences, but included both languages. The memory test can not be used as a control for whether the participants read the sentences, as it was separated into its own experiment, but it still provides valuable input on memory capacity. In addition, easier control questions were added into the SPR-test to control for comprehension while reading, and filler sentences (no

data gathered) were added after each question to use as padding between reading sections.

When developing a new way to assess working memory, it is important to stop and ask whether the test is really testing what we want it to test. And can performance in the memory test be related to the SPR-test?

The memory test tests both verbatim recall and reading span. The participant reads 4, 6, or 8 words, keeps them in short-term memory, and will then read four more words while performing the decision task. This task is an online method of testing working memory, as opposed to Danemans approach 1980 using verbal recall, which was presented in chapter 2.

Using an online method of testing bilingualism may be favorable, as the bilingual will have to rely more on intuition than explicit, learned knowledge. I have not controlled for explicit knowledge of either languages. A recent study by my co-student Julie Matilde Stenhammer Sverreson (Sverreson, Sverreson) showed that there is not necessarily a correlation between explicit knowledge such as a offline test of vocabulary and an online lexical decision task. The factor of linguistic intuition may be extra valuable when parsing GP sentences, as meta-linguistic awareness may impact code-switching.

Testing two languages at once has, to my knowledge, not been done in a working memory test. It adds a new aspect to study bilingualism in working memory. The assessment is done on word-level, as opposed to Daneman's (1980) reading of sentences.

The memory test comes in three versions. The methodological differences between each version has been presented in chapter 4, and stimulus used is listed in Appendix A.

Memory test version A serves as a baseline. It tests working memory in the two languages. Version B proved to be a bit problematic, and had lower task compliance than the other two versions. The idea of testing translations without a match word was to test for translation effect alone, but it was confusing for some participants to see not a direct match, but a translation of a word from the memory set, triggering some to answer yes-match, which may lead to confusion in participants, whether this is correct or not. When asked, all participants except one completing the B-version, solved the task by considering translations as no-match. Version C tests the effect of the combination between match word and translation, to see if there is an increase in RT.

We saw that the mean RT for version C is slower than version A. RT for correct responses in C is 1634 ms, 52 ms quicker than A (1582 ms). For the mean of all RT, including wrong answers is also 50 ms slower in C than in A. In version B, RT is 109 ms slower than A, but this may be explained by the confusion effect.

Testing translation effect alone and translation effect on match words showed an increase in RT, but for version C, there was an increase in correct responses.

6.2 Pilot Study

In this experiment, I found interesting, and perhaps unexpected results, in differences in reading time for a combination of a syntactic phenomena, Garden Paths, and a perhaps more lexical process Code Switching. The results of the pilot study shows that the effect of GP and CS together is not additive. CS does not add to RT of CS+GP sentences, and only slightly affects the comprehension of GP sentences.

We have seen that GP sentences take longer to process generally. Finally we have seen that the syntactic reparsing of Garden Path sentences is marginally affected by Code Switching. It looks like the parsing process considers information at a higher level than the lexical level, and that information integration is not language dependent.

If the reader had to involve two grammar mechanisms for the CS sentences we would expect to have at least additive effects of GP and CS (i.e. a zero or positive rather than a negative interaction factor). It was observed small effects for comprehension of combined GP and CS sentences. The results of the pilot study indicated the need to redo the experiment and better control for variables such as bilingual status, reading proficiency and possible differences in working memory. I learned more about planning the next experiment.

Low task compliance in this experiment showed the need for a retest. The combination of a very challenging comprehension task mixed into the reading time study was not ideal. Mixing two types of tasks in one experiment, and switching between task interchangably is not ideal, especially when measuring RT. As defined by both (Baayen and Milin, 2010) and (Luce, 1991), there are different types of responses within reaction time studies. Response time is to make a conscious decision between two or more answers, and reading time is the time it takes to take in the stimuli before continuing. These two tasks are different processes, and thus the decision making in the comprehension task affects the flow of taking in stimuli in the reading task. Answering the very tricky comprehension question takes the focus away from simply reading to reading to solve a task. Therefore, in experiment 1 and 2, the comprehension questions were followed by a sentence that was discarded and not used for the analysis of GP or CS. Experiment 2 also has a control that the participants are taking in the stimuli, but there is much less switching between tasks, and filler sentences are added after each question, so to not gather data on the sentences after questions and avoid effects due to switching tasks.

6.3 Experiment 1 (or Evaluation of Stimuli)

Experiment 1 was not analyzed in depth due to an unbalance in stimuli for the memory test. As I stated in chapter 4, running the experiment was not a complete waste, as it makes experiment 2 more proved and valid. Other changes to be made in the SPR-test was discovered, and in experiment 2, stimuli was thoroughly controlled. Even though stimuli went through

another round of control, some details slipped.

One cognate made it into the stimuli of the memory test:frog - frosk. Another word that may be problematic is the translational cognate ”kindergarten” that was used in the SPR-test. For the memory test, most words are concrete nouns. I also raised the point of translational homographs in chapter 3. This could have been better controlled for in the final version as well.

I could have been more generous in excluding outliers. I have excluded on a basis of RT. But then, for some of the participants, a lot of data points are removed, but some may still be within the RT boundaries. An exclusion of outliers on participant level could have been carried out, but in this type of exclusion, it is trickier to use the 2 sd from the mean as an exclusion criteria.

In the linear regression model analyses that were carried out, there was little or no variation between blocks of the SPR-test. This confirms no variation between blocks.

Some may find that leaving out commas in GP constructions make them ungrammatical. Commas would disambiguate most of the GP sentences. I argue that using commas in combination with a word-by-word reading experiment is not ideal. In this method, the participants read word by word, and reading a word with a comma following it is not comparable to reading a standalone word. In Norwegian, it is not common to use a comma after noun phrases in itself. Example 38 is not a correct way of using comma in Norwegian, although it disambiguates the sentence by making ”hadde elsket” impossible.

(38) Hunden jeg hadde, elsket bein.
The dog I had, loved bones.

This is mostly a question of norms, and not about language processing. It may be looked upon as noise in the stimuli. The human language processor is able to process much more local ambiguity without problems. The main interest of GP sentences is to investigate how language is processed, not what is correct in a language or not.

6.4 Experiment 2

The results of experiment 2 correspond to the results of the pilot study, only that the results of experiment 2 are more proofed and thus more safe/valid.

6.4.1 Comparing ”endpos” to ”point of GP/CS”

In the analysis, we analyzed both the end of the sentence and the point at which the GP and CS occurs (or doesn’t occur). See section 4.5.5 for procedure on tagging files for analysis.

Why is it relevant to look at both? These two different ways of analyzing the data may show different things. It shows where the effect happens, and gives a ground for comparison.

Table 6.1: Memory Test: Correct responses by set length
4-set: 408 correct
6-set: 380 correct
8-set: 366 correct
Sum: 1154

These two different analyses yield results that correspond. There is a stronger effect of both GP and CS at the point where they occur.

End effects:

RT = 482.1 + 17.8(GP)* + 19.9(CS)* - 15.4(GP + CS)

Point of GP/CS:

RT = 459.2 + 25.7(GP)** + 35.3(CS)** - 40.9(GP + CS)*

For the point of GP/CS, GP adds a significant extra reading time of 25.7 ms at the point of GP, CS adds a significant extra reading time of 35.3 ms at the point of CS, and GP and CS combined subtracts 40.9 ms to expected reading time, also significant.

In the sentence end, effects are smaller, with an added reading time of 17.8 ms for GP, 19.9 for CS, and interaction -15.4 for the antepenultimate, penultimate and ultimate sentence word. This was not significant when applying the ANOVA.

Both the GP and the CS effect are clearer at the point where they occur. The effect at the point that separates the baseline from GP/CS is larger than at the end of the sentence.

6.4.2 Memory Test

As Sternberg (1966) stated, both yes and no answers should increase in RT for difficulty. His results also showed that no-answers take slightly longer than yes-answers.

The figure 6.1 was given in the results chapter, is here reprinted for ease. Comparing the figure 6.1 showing results in RT per language and set length to the table 6.1 showing correct responses per set length.

The figure 6.1 shows reaction times for each of the set lengths for each of the languages. We see a difference in languages, and in set lengths. The main difference is set length 8 for English. It is also the only one that is significantly different from the others. Oddly, for Norwegian it seems that the 6 length is slower than 8, but the variance is high, and the difference is not significant. The results imply that there are extra resource demands for English set length 8. There is a significant effect of match language English and set length 8:

Regression: (ReactionTime-length*matchlang +((matchlang*length)|Participant))

ANOVA: $F(2, 23.2) = 5.0$, $p<0.01$*

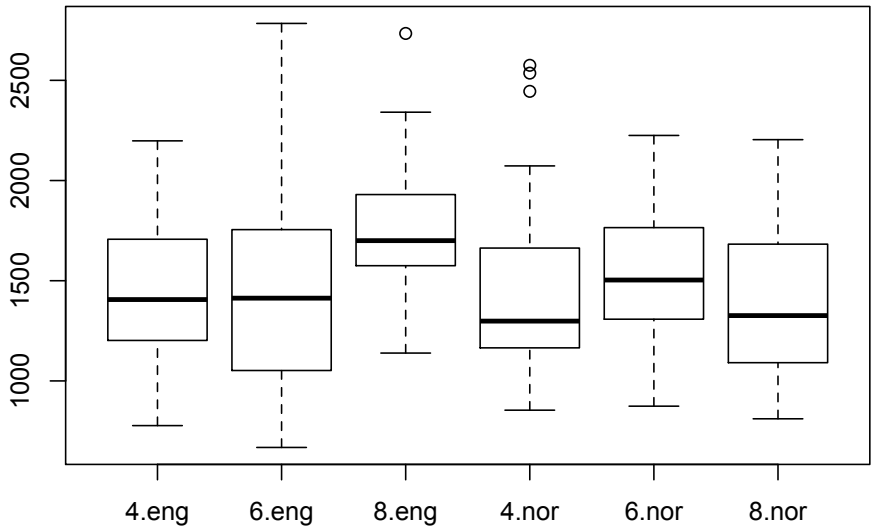

Figure 6.1: Memory Test: Boxplot of length of memory set and language for A-set

Comparing both correct responses per language and RT per language is valuable. There is an increased RT for length 8 English, and also less correct responses. The increase of RT, as shown in figure 6.1 shows a significant increase at set length 8 only for English.

There is a difference in correct responses for set length, and for language the differences are varying. For match language and set length, there was varying amounts of correct responses. For memory set 6, there were 53 more correct responses for English, whereas for memory set 8, there were 37 more correct for Norwegian.

One study by Marian and Neisser (2000) showed in an experimental study that events are better recalled if the language used to recall them is the language in which the event took place. They called it ”language-dependent recall”. The results of the study may explain why translations in memory version B and C do not facilitate quicker response times. The mean RTs for correct responses in blocks A B and C were 1582 ms, 1691 ms, and 1634 ms. Another explanation may be that it adds to the cognitive load of the task to see two matches. On the other hand, two congruent signals pointing in the same direction should be reacted to faster, not slower. This needs to be further validated, as there were many missing data points in version B, thus giving it less validity. Translations on match words lead to more correct responses. Translations do not facilitate quicker RT, but may facilitate correct responses.

6.5 Results of Pilot Study compared to Experiment 2

In the pilot study, we saw a highly significant effect of GP, and a near-significant effect of CS. In experiment 2, we saw a significant effect both at the point of GP and at the point of CS. One explanation to why the results are different is the method of SPR used. In the pilot study, sentences were presented in halves, which may induce the effect of GP because it presents the reader with a clause, and then in the other half one sees that this clause is indeed not a clause, as in example 39.

(39) Mannen som hører stemmer | en fin saksofon.
The man who hears (voices|tunes) | a nice saxophone.

(40) Mannen som hører stemmer | plays saxphone well.
The man who hears (voices|tunes) ...

For CS sentences presented in halves, it may lessen the effect of the CS because there is a clearer separation between the languages. One half is solely in Norwegian, and the other half solely in English. More importantly, it is soon expected that half of the sentences switch language. This is not the case in word-by-word, since most words are followed by a word in the same language. As we see in example 40, there is a clearer separation between the languages, as opposed to reading word by word where the change can happen at many different points. The effect is integrated as a reading effect for all the words in the last half.

(41) Forfatteren / mente / det / samme / var / sagt / tidligere.
The writer / gave opinion / the / same / was / said / earlier.

(42) Forfatteren / mente / at / everyone / should / read / the / book.
The writer / gave opinion / at / everyone / should / read / the / book.

In experiment 2, moving window word-by-word self paced reading was used. The moving window reveals only one word at a time, and the position of the word is visualized from the start. This is beneficial and also of higher ecological validity, as it does not present preferences of the sentence and is a more neutral method of testing. As we see in example 41, the reader is more free to read and analyze here, as there is no priming of structure through forced separations of clauses. In example 42, the reader does not know at which word in the sentence the CS may occur or if it occurs at all, as opposed to in the pilot study where the CS occurs in the second half.

Common for both experiments is that GP and CS does not yield an additive effect, suggesting that the processes of top-down and bottom-up information are not separable. The results of the two experiments vary slightly, as expected when using different methods, but the results are not contradictory.

That the effects are decreasing is in line with results from reproducing studies across all fields of science. The term verbal overshadowing was introduced in a series of reproduced studies that showed a decrease in significant results (Lehrer, 2010).

In my results, the decreasing effects may be explained by the change in methods. A more sound and ecologically valid method has been used for experiment 2 compared to the pilot study. Task compliance was much higher in experiment 2.

6.6 Aspects of Garden Paths

I have chosen a different approach to Gibson in studying bottom-up and top-down information. Gibson (2006) also used SPR, but limited the scope of his ambiguous sentences to ambiguities using the word ”that”. The benefits of using a narrower scope of ambiguous sentences is that they are more comparable, while my results are based on a broader selection of structures, and is thus more general.

Gibson (2006) examines one environment to test whether people are sensitive to the lexical category-frequency distributions of each word, independent of context (Gibson, 2006). He gives an example:

(43) a) The lawyer for that skilled surgeon asked for a raise.
b) The lawyer for those skilled surgeons asked for a raise.

The ambiguity of example 43 lies in reading ”that” as either a determiner or preposition. The prediction is that A is more difficult than B, which is supported by Gibson’s results.

I chose to use CS as the test of bottom-up lexical information. While Gibson’s 2006 study is monolingually English (whether the participants are monolingual or not is not specified since the study uses only one language), I have used Norwegian and English and bilinguals in those languages.

I have used a range of structures that fit within the definition of the GP. The chances are low that a participant will miss all the different structures that the GP sentences are. An analysis of RTs within the group of GP sentences was presented for the pilot study in chapter 3.

There is a variation within the group of GPs, therefore I found it natural to maintain a variation within the group of baseline sentences. By variation, I mean that there is not only one sentence structure of GP that is tested, and so this should also be true for the baseline.

There is no standard structure for what the baseline is, it is only a ”normal” version of the GP sentence presented in another block. There is a variation within the group of GPs, therefore I found it natural to maintain a variation within the group of baseline sentences. By variation, I mean that there is not only one sentence structure of GP that is tested, and so this should also be true for the baseline. When analyzing the data in the lmer-model, I found no effects per block. This indicates that there was no effect of what block the sentence was presented in, and shows that the variations across sentence blocks was not affecting the

model. Using a variation of structures, makes the effect more externally valid. This is not the case in the study by (Gibson, 2006), where only one construction is tested.

(44) Bonden / gir / dyrene / mat / og / tørt / høy.
The farmer / gives / animals / food / and / dry / hay.

In sentence 44 we see a normal sentence. The examples 44 and 45 share the first three words of the sentence. Here, the baseline contains 7 words and the GP 8, but this is balanced out when all sentences are taken into account.

(45) Bonden / gir / dyrene / ull / blir / klippet / av / vann.
The farmer / gives / animals / wool / is / cut / of / water.

The definition of GP leaves room for interpretation, so to be able to compare it to the broad span of GPs, it is natural that there is variance within the baseline group as well. Also, by providing different sentence structures for the baseline, the results are more likely to generalize. There was a significant effect of GP, even though there are many types of GP structures, and also many types of baseline structures. Therefore, in the test, it will be less predictable what is a GP or a baseline, as the participant does not know what to look for. The beginnings of the sentences (before point of GP/CS) are the same across blocks, and thus the effect will not be revealed until the point of GP/CS.

Gibson's 2006 results support his theory of top-down/bottom-up. The relative frequencies of lexical entries for a word are tabulated independent of context. Gibson uses the same method as I used for my experiment 1 and 2: self-paced word-by-word moving window. Like me, he also used simple yes/no-questions for information from the sentence, with no feedback. He controlled for word length of "that" versus "those" with a regression equation. He found a significant increase in reading pace for positions 4-6 (that|those skilled surgeon(s)) (Gibson, 2006). I also found effects of change in bottom-up information through CS, and top-down through GP. I have chosen to test Gibson's (2006) theory differently from Gibson himself. Where Gibson only tested the effect of one structure, I have tested the theory in a broader and also bilingual sense. My results support his theory that both bottom-up and top-down information is used in processing.Gibson's (2006) theory on syntactic ambiguity resolving proposes that people use syntactic (top-down), and lexical (bottom-up) information to resolve ambiguity. There is a significant load of both top-down and bottom-up processing, but when both are challenged, there is no additive effect.

6.7 High Spanners vs Low Spanners

Just and Carpenter 1992 coined the terms high spanners and low spanners. They claim that only high span individuals have sufficient working memory capacity to use contextual information to disambiguate syntactic ambiguities. For my experiment, maybe this will show in

a correlation between high spanner - low RT on GP+CS. Also, low spanners may show an additive effect of GP+CS?

The size of the experiment limits the conclusions we can draw based on it. The test comprises 24 data points per participant, with three set lengths. A future experiment should comprise a new set length of 10 words in the memory set, to better be able to separate high spanners and low spanners. As the test stands now, most participants score quite well. The memory sets of 4 and 6 do not lay the ground for a separation measure, it is at the 8-set we see more clearly the effect of set length and language.

Due to the limited scope of the experiment, it is hard to draw conclusions on the differences in high spanners and low spanners.

Miller (1956) states that the average person can keep 7 words in his or her short-term memory at a time, with a variance of +/- 2. So the set length of 8, will only separate the low spanners, but not allow the high spanners to be separated as easily. Including a set of 10, or even 12, will more clearly separate high spanners from low spanners. It can also more clearly test the effect of L2 in longer, more demanding sets. I found no significant differences in language for set lengths 4 and 6, but maybe as the task grows more challenging, there will be a larger effect of language. This has been indicated through a significant effect of set length 8 and English, but can be further validated through longer set lengths.

There is a clear and definite limit to the accuracy with which we can identify absolutely the magnitude of a unidimensional stimulus variable. Miller (1956) maintains that for unidimensional judgments, this span is usually somewhere in the neighborhood of seven. But what about testing two langugages, does it count as two dimensions? Is there an extra load of activating L2?

My results indicate that there is an added effect of activating two languages in working memory, as shown by a significant effect of match language English on set length 8.

6.8 Interpreting Correlations

I found no correlations between the results in RT between the memory test and the SPR-test. H0 was not discarded: There is no correlation between performance on SPR-test and memory test. Correct responses and reaction times in memory test do not correlate with reaction times on SPR-test. An increase or decrease in RT per participant, do not correspond to an increase or decrease in RT in the SPR-test for that participant. So we know that there is no correlation, but what can that imply?

The two tests are different tasks. The memory test is a decision task, the SPR-test is a continuous task. As we saw in the pilot study, there was a low task compliance when two tasks were combined into one experiment, but maybe it means that two different types of tasks represent different processes, and are thus not comparable.

Some literature, such as (Just and Carpenter, 1992), claim that there is a correlation be-

tween working memory and reading. As we saw in the previous section 6.7, the memory test does not give enough data points to clearly separate high spanners from low spanners. Caplan and Waters (1999) claim there are different processes underlying working memory and reading, and that they are therefore not comparable. Caplan and Waters (1999) claim there are different linguistic processes of working memory. One process involves the on-line, unconscious, psycholinguistic processes of comprehension, and the other involves controlled verbally mediated task. My research support Caplan and Waters' (1999) claim.

Not only are the two tests different tasks, they are testing different things. The SPR-test tests for reading flow, which does not demand a conscious decision. The memory test requires the participant to make a conscious decision, and applying explicit knowledge. So therefore, no correlation between the tests can imply that the unconscious effort of reading, and even bilingual reading, is not depending on explicit word recognition or language knowledge.

6.9 The Discount of Bilingualism

We have now seen the results of the study, analyzed it for significance, and then discussed it. I found evidence that the combination of GP and CS is significantly faster than predicted by an additive effect. Both processes contribute information that together makes the task easier. But what do the results imply?

Across the fields of second language acquisition, bilingualism and psycholinguistics, there is an ongoing discussion of the price of bilingualism. Many articles, both scientific ones based on studies and evidence, and articles aimed at a broader audience without the same scientific backing make claims about the price of being bilingual. Typical arguments for bilingualism claim that being bilingual has benefits other than the direct linguistic benefit, such as better working memory or greater meta-linguistic awareness (Baker, 2006). Research has also shown that bilinguals are better at ignoring distractions (Highby et al., 2015) Arguments against bilingualism show to slower development of one of the languages in some children.

Field (2011) points out that to know a word is to know its grammatical relations. Grammatical relations may be stored in the bilingual's mental lexicon on a word level, such that when CS is combined with GP, the syntactic ambiguity is not stored as language-specific, and therefore will not add more of a cognitive load.

In chapter 3 we saw two sentences that were equal on structure level, but one was a GP and the other was not, showing that lexical information can guide sentence structure. This may be an explanation to why there is no added effect of CS on GP sentences, as the load of lexical information is already induced.

In my experiments, I have studied the cost of GP and the cost of CS, and found that when both GP and CS are combined, there is a discount in cognitive load (as measured in RT). The effect of GP and CS is not additive. Activating a second language (in form of CS) may facil-

itate alternative parses in the reading process. This may be an explanation to why the effect is not additive. Just as monolinguals possess intuitions about what constitutes well-formed utterances in their native language, bilinguals have the capacity to differentiate ill-formed from grammatical patterns of CS (Gullberg et al., 2009). The metalinguistic awareness that CS or macaronic language brings may make it easier to read an ambiguous sentence.

Baker 2006 showed in a study that bilingual children appeared to have greater metalinguistic awareness, and a more analytical view of language compared to monolinguals, and Baker 2006 suggested that the ability to control two languages enabled the bilingual children to perform better on tasks such as counting words in a sentence. This may then be related to CSing, and the metalinguistic awareness of seeing other parses of the GP.

6.10 Further Research

One thing that will be very interesting to do is to extend the length of sets for the memory test. It would be very interesting to add a memory set of 10 words. According to Miller (1956), the amount of words a person can keep in short-term memory is 7 +/- 2. This corresponds to the results we have seen, where the overall difficulty starts at the set of 8 in English. Adding to the length of the test will make it possible to use as a stand-alone test to control for balanced bilingualism in working memory, and not just as a correlation factor for this experiment. Including a set of 10, or even 12, will more clearly separate high spanners from low spanners. It can also more clearly test the effect of L2 in longer, more demanding sets.

Furthermore, the test could be used on a wider audience, not only university students, to make the results more generally valid. The memory test can also be used in other settings than to correlate with the SPR-test as we have done here. An extended version of the test can be used on its own to assess a persons bilingual balance and working memory span. Running an extended version of the test will be a good indicator of bilingual balance in working memory, and there is, to my knowledge, currently no test that does this. The test can then be used on different level for proficiency. Of special interest is the difficulty for beginner learners.

The SPR-test could also be extended in amount of sentences used. It would also be interesting to run this experiment with eye-tracking equipment. We have now seen robust results in the SPR-test, will they be confirmed again through the use of eye-tracking?

I could always have chosen other methods or analyses. One interesting method of word-by-word reading is Kizach's (2013) word-decision making in the process called G-maze. Words are presented, and the participant will have to choose between two words to continue a sentence. This is good for GP sentences, because when presented with two alternatives, a word that creates a GP or a word that makes the sentence ungrammatical, the participant, if he or she sees the GP-effect, will choose the word for GP, and if participants do not see the GP effect, there will be a higher error rate at this point. CS can be implemented in the same way, testing for grammaticality acceptment of intrasentential CS. This would solve the

verification of reading.

Testing grammaticality in CS is very interesting. I have pointed at different factors that make written CS challenging. In the same way as I in this thesis have looked at the influence of CS on syntactic structure, it would be of interest to take a closer look at different types of grammaticality issues in CS structures.

Chapter 7

Conclusion

Limiting myself to the scope of this thesis has been challenging. I could easily have spent another year testing an extended version of the memory test or expanding the experiment, but that will have to wait for future research.

My main findings are that there is a significant increase in reading pace for GP sentences and CS sentences, as proven by both the pilot study and replicated in experiment 2. When both CS and GP are present in the same sentence, there is no additive effect. The main effect of both CS and GP are immediate, significance was found at the point of GP/CS, and not at the sentence end.

I have detected an imbalance in working memory capacity between L1 and L2. I found a significant effect of set length 8 and match language English in the memory test. Translations slowed RTs, but slightly facilitated correct responses on match words. Approaching the working memory limit shows differential effects in L1 and L2 being slower and less accurate. Distribution on a participant level varies, and shows no correlation to performance in SPR-test.

No correlation was found between the RTs of the memory test and the SPR-test. There is no correlation between reading pace of the different sentence types in the SPR-test and match language of the memory test.

There are both challenges and risks of developing a new test of bilingualism in working memory. In the end it proved rewarding, and I hope in the future to extend the length of the experiment to make it stand on it's own as it's own measure of working memory and bilingual lexical access, not just as a control factor for this experiment. Bever (1970) approached language as a conceptual and communicative system which recruits various kinds of human behavior, but is not manifested in a particular form of language behavior. Therefore it is important to apply different methods in testing different methods of language. My memory test is a new way of assessing bilingual working memory, and although it can be further improved, it provides a valuable addition.

By showing that there is no additive effect of syntactic processing of GP (represented by GP) and lexical access of CSing, I argue that the syntactic parsing and lexical access

are not separate linguistic processes. Code-switching does not add processing difficulty to syntactic ambiguity for Garden Path, but may help us see other alternative parses. Activating a second language (in form of CS) may facilitate alternative parses in the reading process. Grammatical relations may be stored on a word level, such that when CS is combined with GP, the syntactic ambiguity is not stored as language-specific, and therefore will not add more of a cognitive load.

As the results have been presented and then discussed, Bever's question of how a person arrives at internal linguistic knowledge from external output sequences still remain unanswered. But I have provided new data and results on how different output sequences affect internal linguistic processes.

Appendix A

Memory task as presented in Experiment 2

The search set is the same in both versions. In the target set, the word that is switched for a translation is *emphasized* and the translation is in parentheses. Version A has no translations, version B has translation in all of the lines, version C has translations only in lines with a match (related).

Lead-ins (same in all versions)
bryst step trinn watch CALL BUNN WOOL BRYST
pensum regn mouse ticket grammar skap TARGET STRIKK BLYANT ABUSE
ran sko voice flue granat screen face hour FLUE LICK MAPPE PUTE

4 related
bille fork smør oat BILLE SEED LANGUAGE *FJELL* (HAVRE)
button peace tau kinn *DRUE* PEACE SQUARE BØLGE (KNAPP)
month crown hund dialekt SAUSAGE *JENTE* HUND ROOF (ACCENT)
blomst chair beer tallerken APPELSIN FARMER *PLAKAT* BEER (STOL)

4 unrelated
horse sykkel sweater gryte *INGEFÆR* JERN CUCUMBER TISSUE (HEST)
kjeks nail blad sheep BIL FRAME PANTS *KURV* (SAU)
søppel table fugl poem *CITY* FLASKE REV LIBRARY (BIRD)
wheel mat tegning candle *KASSE* SPEIL WAIST PILLOW (FOOD)

6 related
stairs smykke lapp coat tårn feeling STAIRS BRILLE *RYGG* GLUE (FRAKK)
building gardin crumble seed elg prøve *FARM* GARDIN HYLLE PUMPKIN (MOOSE)
lomme mother kjole pinne foam leaf *FLAGG* LEAF GOAL FEIL (MOR)

owl path map sal frosk veps KEY STILK *PART* VEPS (FROG)

6 unrelated

lås bokstav monkey shower kite brev SAMFUNN DOUGH SØSKEN *TEAM* (LETTER)
tooth skinke barley vask klokke child TIMIAN YEAST OST SOUTH
skrue ear science kålrot bue cloud BJEFF *LUE* PEBBLE BRANCH (øre)
candle carrot flette journey butikk fløte STEIN BRAIN FURNITURE *MOSE* (REISE)

8 related

jord home bær doll fylke bride hanske ant FYLKE *CARD* ETASJE TORCH (EARTH)
bone record humle picture sag promp seng krok SIGN KROK STØVEL *MEADOW* (SAW)
høre ghost suppe map divorce pregnant seng kirurg SALMON SÅR GHOST *RIVE* (KART)
janitor porridge dåp tastatur woman frisør evening frokost IDRETT COUPLE *BYGG* PORRIDGE (DAME)

8 unrelated

ære undulat ødemark vinke eagle autumn lumber toe BLÅSE BULL *UVÆR* BRUSH (ØRN)
vår hedge attack enviroment needle nærsynt morgen kile LOAN CHAPTER LILJE *PADDE* (HEKK)
property kill nebb page løk garn mercy brus SUKK *PROTECTION* BRYLLUP JUDGE (ONION)
dråpe panne wolf happiness cancer grav barbeque brød GREED *FLYKNING* NECK MARSVIN (ULV)

Appendix B

Sentences as presented in Experiment 2

Training sentences and questions are the same across all blocks.

Training sentences:

1. Ole er snill mot den oransje katten sin.
2. Andre barn leker aldri outside when it rains.
3. Zoologen gir fugler egg blir tatt fra mat.
4. Bøker med gule omslag are more frequently read.

Questions:

1. GP Anne slår hunden katten plaget.
Inneholdt setningen du nettopp leste ordet «hunden»?
CS Sykkelen i skuret is newer than mine.

2. BL Kaffekoppen er full av deilig kaffe.
Inneholdt setningen du nettopp leste ordet «sko»?
GPCS Julie puttet dropset i munnen on the table.

3. GP Jon gir barn trygghet blir tatt fra hjelp.
Inneholdt setningen du nettopp leste ordet «barn»?
CS Mine rosa sokker have holes on both heels.

4. CSGP Den flinke mannen fikk beskjeden pay for his effort.
Inneholdt setningen du nettopp leste ordet «brød»?
BL Masterstudenten på lesesalen gråt i stillhet.

Block 1:

GP 1. Mannen som hører stemmer en fin saksofon.
BL 2. Kari ga barnet hunden og en leke.

GPCS 3. Det beste barnet visste was red apples.
CS 4. Den andre kokken laget soup with herbs.
GP 5. Det siste barnet ville spise var pannekaker.
GP 6. Det minste barnet kunne gjøre var å rydde.
BL 7. Læreren presenterte gutten for den nye klassen.
CS 8. Per gir fisken middag and other food.
BL 9. Politikeren i salen er alltid til stede.
GPCS 10. Hunden krøllet sammen on the carpet sleeps.
GPCS 11. Forfatteren mente det samme had already been said.
CS 12. Bonden gir dyrene food and dry hay.

Block 2:

BL 1. Mannen som hører stemmer spiller fint saksofon.
GPCS 2. Kari ga barnet hunden licked a toy.
CS 3. Det beste barnet visste the correct answer.
GP 4. Den andre kokken laget spiste vi med urter.
BL 5. Det siste barnet ville spise søte pannekaker.
BL 6. Det minste barnet kunne gjøre seg forstått.
GPCS 7. Læreren presenterte gutten for the girl asked.
GP 8. Per gir fisken middag lages av mat.
GP 9. Politikeren i salen av mykt lær red avgårde.
CS 10. Hunden krøllet sammen the freshly ironed tablecloth.
CS 11. Forfatteren mente at everyone should read the book.
GPCS 12. Bonden gir dyrene wool is cut from water.

Block 3:

GPCS 1. Mannen som hører stemmer a nice saxophone.
CS 2. Kari ga barnet hunden and a toy.
GP 3. Det beste barnet visste svaret på spørsmålet.
BL 4. Den andre kokken laget saus med urter.
GPCS 5. Det siste barnet ville eat sweet pancakes.
GPCS 6. Det minste barnet kunne gjøre was clean up.
CS 7. Læreren presenterte gutten to the new class.
BL 8. Per gir fisken middag og annen god mat.
CS 9. Politikeren i salen is always present during meetings.
GP 10. Hunden krøllet sammen på teppet sover.
GP 11. Forfatteren mente det samme var sagt tidligere.
BL 12. Bonden gir dyrene mat og tørt høy.

Block 4:

CS 1. Mannen som hører stemmer plays saxophone well.
GP 2. Kari ga barnet hunden slikket en leke.
BL 3. Det beste barnet visste var å lese.
GPCS 4. Den andre kokken laget we ate with herbs.
CS 5. Det siste barnet i kindergarten wanted pancakes.
CS 6. Det minste barnet kunne make itself understood.
GP 7. Læreren presenterte gutten for jenta ba om det.
GPCS 8. Per gir fisken middag er made from food.
GPCS 9. Politikeren i salen made of leather rode away.
BL 10. Hunden krøllet sammen den fine nystrøkne duken.
BL 11. Forfatteren mente at alle burde lese boken.
GP 12. Bonden gir dyrene ull blir klippet av vann.

Appendix C

Memory Test Box plots of RT per participant

Box plots for each of the three versions of the memory test. Box plots show reaction time per language per participant, and may be used as an indicator of balancedness per participant. Participants with approximately equal RT for both languages are balanced. Participants with different RT for each language are favoring one language over the other. An example of a balanced participant is participant 37. An example of an unbalanced participant is participant 26.

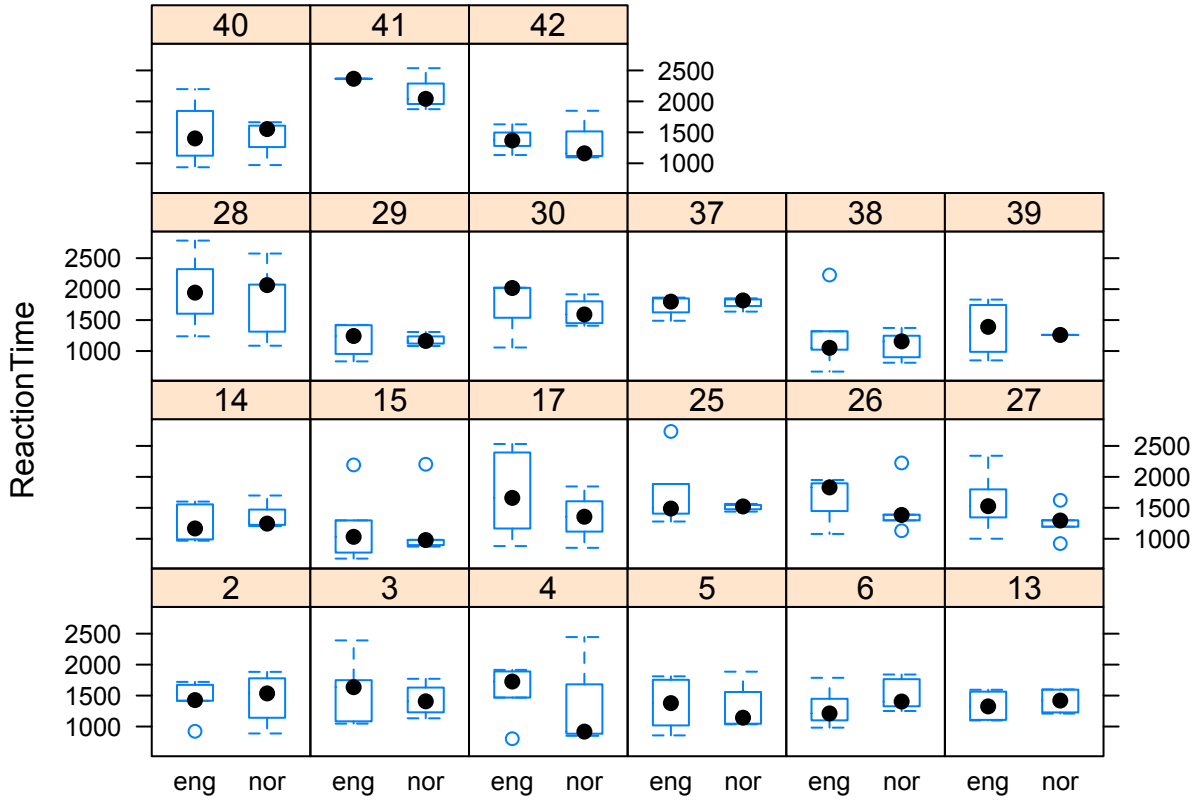

Figure C.1: Memory Test: Box plot of Reaction Time per Participant for each language in A-set

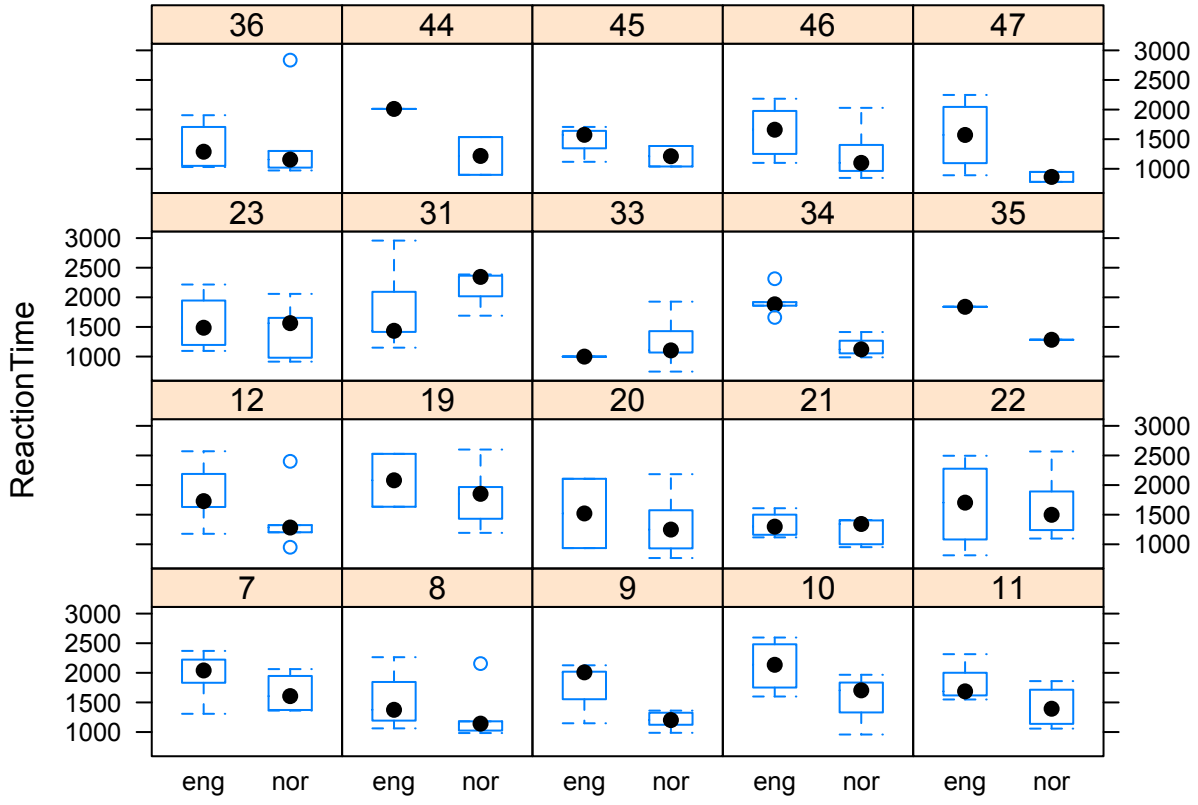

Figure C.2: Memory Test: Box plot of Reaction Time per Participant for each language in B-set

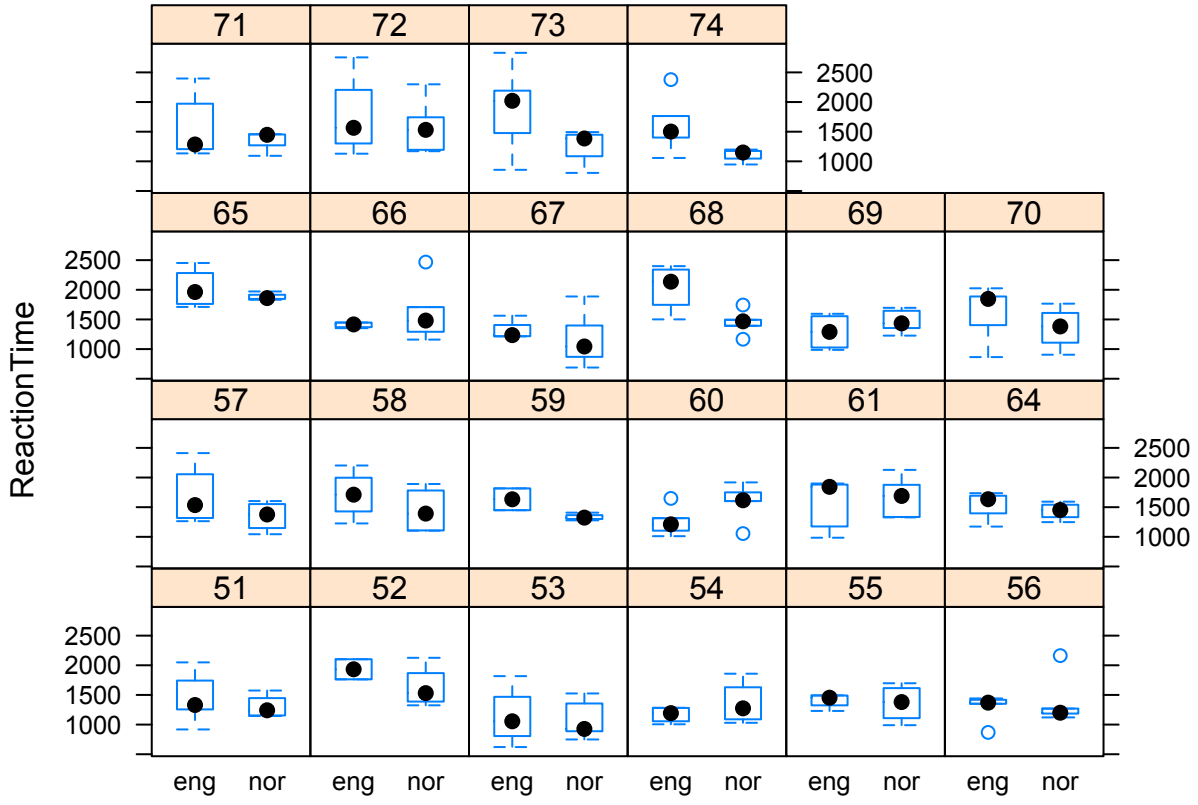

Figure C.3: Memory Test: Box plot of Reaction Time per Participant for each language in C-set

Bibliography

(2009, February). Makaronisk poesi. Store Norske Leksikon. 10, 11

(2016, april). Cedrus rb-series. 39

Baayen, R. H. (2009). *Analyzing Linguistic Data - A practical introduction to statistics using R*. Cambridge University Press. 41, 49

Baayen, R. H. (2013). Multivariate statistics. In R. Podesva and D. Sharma (Eds.), *Research Methods in Linguistics*, Chapter 16, pp. 337–372. Cambridge University Press. 49

Baayen, R. H. and P. Milin (2010). Analyzing reaction times. *International Journal of Psychological Research, 2010 3*(2). 32, 33, 46, 50, 78

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences 4*(11), 417–423. 20

Baddeley, A. and G. Hitch (1974). Working memory. *The psychology of Learning and Motivation, 8*, 47–90. 20

Baker, C. (2006). *Foundations of bilingual education and bilingualism*. Multilingual Matters. 2, 11, 86, 87

Bauvillain, C. and J. Grainger (1987). Accessing interlexical homographs. *Journal of Memory and Language 26*, 658–672. 28

Beatie, B. (1967). Macaronic poetry in the carmina burana. *Vivarium 5*(1), 16–24. 10

Bever, T. (1970). Language down the garden path - the cognitive and biological basis for linguistic structures. *Cognition and the development of Language*, 279–362. 1, 12, 13, 23, 24, 89

Bresnan, J. (Ed.) (1982). *The Mental Representation of Grammatical Relations*. The Massachusettts Institute of Technology. 14

Bullock, B. E. and A. J. Toribio (Eds.) (2009). *Linguistic Code-switching*. Cambridge University Press. 2, 5, 9, 10, 12, 48

Caplan, D. and G. S. Waters (1999). Verbal working memory and sentence comprehension. *Behavioural and Brain Sciences 22*, 77–126. 17, 21, 86

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press. 1

Colman, A. (Ed.) (2008). *The Oxford Dictionary of Psychology* (4th ed.). Oxford University Press. 5, 20

Corbin, L. and J. Marquer (2009). Individual differences in sternberg's memory scanning task. *Acta Psychologica 131*. vi, viii, 3, 22

Daneman, M. and P. A. Carpenter (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behaviour 19*, 450–66. 2, 17, 20, 21, 33, 77

Eckert, P. (2013). Ethics in linguistic research. In R. Podesva and D. Sharma (Eds.), *Research Methods in Linguistics*, Chapter 2, pp. 11–26. Cambridge University Press. 48

Field, F. W. (2011). *Key concepts in Bilingualism*. Palgrave key concepts. 6, 10, 11, 12, 16, 19, 86

Frazier, L. (1979). *On Comprehending Sentences: Syntactic Parsing Strategies*. Ph. D. thesis, University of Conneticut. 2, 12, 13, 14

Gathercole, S. (2009). Working memory and language. In G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics*, Chapter 46, pp. 757–770. Oxford University Press. 5, 17, 20, 50, 76

Gernsbacher, M. A. (1994). *Handbook of Psycholinguistics*. Academic Press. 17, 40, 41

Gibson, E. A. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition 68*, 1–76. v, vii, 12, 15

Gibson, E. A. (2006, April). The interaction of top–down and bottom–up statistics in the resolution of syntactic category ambiguity □. *Journal of Memory and Language 54*(3), 363–388. v, vii, 2, 12, 15, 16, 31, 47, 75, 83, 84

Gries, S. (2013). Basic significance testing. In R. Podesva and D. Sharma (Eds.), *Research Methods in Linguistics*, Chapter 15, pp. 316–336. Cambridge University Press. 49

Grosjean, F. and P. Li (2012). *The Psycholinguistics of Bilingualism*. CWiley-Blackwell. 2, 6, 10, 18

Gullberg, M., P. Indefrey, and P. Muysken (2009). Research techniques for the study of code-switching. In B. Bullock and A. J. Toribio (Eds.), *Linguistic Code-Switching*, Chapter 2, pp. 21–39. Cambridge University Press. 11, 12, 87

Hernandez, A., E. Fernandez, and N. Aznar-Bese (2009). Bilingual sentence processing. In G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics*, Chapter 22, pp. 371–384. Oxford University Press. 20

Highby, E., S. Donnely, and J. Yoon (2015). The effect of second language proficiency on inhibitory control: An ex-gaussian analysis. *Conference paper at CUNY*. 86

Highby, E., I. Vargas, S. P. an Wendy Ramirez, E. Fernandez, V. Schafer, and L. Obler (2015). Shared syntax for bilinguals extends to language-specific constructions. *Conference Paper at CUNY*. 29

Jared, D. (2015). Literacy and literacy development in bilinguals. In A. Pollatsek and R. Treiman (Eds.), *The Oxford Handbook of Reading*, Chapter 12, pp. 165–182. Oxford University Press. 47

Jurafsky, D. and J. H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.). Upper Saddle River, NJ, USA: Pearson International Edition. 24, 25, 26

Just, M. A. and P. A. Carpenter (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* (99), 122–149. v, vii, viii, 2, 17, 20, 21, 33, 84, 85

Just, M. A. and J. King (1991). Individual differences in syntactic processing: the role of working memory. *Journal of memory and language 30*, 580–602. 33

Kaiser, E. (2013). Experimental paradigms in psycholinguistics. In R. Podesva and D. Sharma (Eds.), *Research Methods in Linguistics*, Chapter 8, pp. 135–168. Cambridge University Press. 12, 17, 32, 35, 42

Kempen, G. (1996). Computational models of syntactic processing in language comprehension. In T. Dijkstra and K. de Smedt (Eds.), *Computational Psycholinguistiscs*, Chapter 8, pp. 192–220. Taylor & Francis, London. 12, 17

Kizach, J., A. M. Nyvad, and K. R. Christensen (2013). Structure before meaning: Sentence processing, plausibility, and subcategorization. 87

Lanza, E. (2008). Selecting individuals, groups and sites. In L. Wei and M. G. Moyer (Eds.), *Research Methodsin Bilingualism and Multilingualism*, Chapter 5, pp. 73–87. Blackwell Publishing. 46, 48

Lehrer, J. (2010). The truth wears off - is there something wrong with the scientific method? *Annals of Science - The New Yorker* (December 13). 82

Luce, R. D. (1991). *Response Times: Their Role in Infering Mental Organization*. Oxford University Press. 32, 41, 78

Luk, G. and E. Bialystok (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology 25*(5), 605–621. 2, 3, 18, 19

MacDonald, M. and M. Christensen (2002). Reassessing working memory: Comment on just and carpenter (1992)and waters and caplan (1996). *Psychological Review 109*(1), 35–54. 21

MacDonald, M. C., M. A. Just, and P. A. .Carpenter (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology 24*, 56–98. 13, 14, 17, 18, 33

Marian, V. and U. Neisser (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology 129*(3), 361–368. 19, 44, 81

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63*, 81–97. 21, 22, 85, 87

Moyer, M. G. (2008). Research as practice: Linking theory, method, and data. In L. Wei and M. G. Moyer (Eds.), *Research Methodsin Bilingualism and Multilingualism*, Chapter 2, pp. 18–31. Blackwell Publishing. 27, 46, 49

Niv, M. (1993). *A Computational Model of Syntactic Processing: Ambiguity Resolution from Interpretation*. Ph. D. thesis, University of Pennsylvania. 12, 14, 15, 20

Pinker, S. (2014). *The Sense of Style - The Thinking Person's Guide to Writing in the 21st Century*. Allen Lane. 76

Podesva, R. J. and D. Sharma (2013). *Research Methods in Linguistics*. Cambridge University Press. 32, 41

Potter, M. C., K. So, B. Eckhardt, and L. Feldman (1984). Lexical and conceptual representation in beginning and proficient bilinguals. *Journal of Verbal Learning and Verbal Behaviour 23*(1), 23–38. 11

Rayner, K. and L. Frazier (1987). Parsing temporarily ambiguous sentences. *The Quarterly Journal of Experimental Psychology 39*(4), 657–673. 12, 13, 14

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc. 46

Sanz, M., I. Laka, and M. K. Tanenhaus (Eds.) (2013). *Language Down the Garden Path - The Cognitive and Biological Basis for Linguistic Structures*. Oxford University Press. 1, 13, 23, 75

Staub, A. (2015). Reading sentences: Syntactic parsing and semantic interpretation. In A. Pollatsek and R. Treiman (Eds.), *The Oxford Handbook of Reading*, Chapter 14, pp. 202–216. Oxford University Press. 12, 13, 16

Sternberg, S. (1966). High-speed scanning in human memory. *Science 153*, 652–654. vi, viii, 3, 22, 43, 58, 67, 80

Sternberg, S. (1969). Memory scanning: Mental processes revealed by reaction time experiments. *American Scientist 57*(4), 421–457. 22

Sverreson, J. M. S. Priming adult beginner learners - a study of cross-linguistic lexical priming in german and spanish learners of norwegian. Masters Thesis of Linguistics at University of Bergen, june 2016. 77

Traxler, M. (2012). *Introduction to psycholinguistics - Understanding Language Science*. Wiley-Blackwell. 14, 24

van Gompel, M., M. Pickering, and M. Traxler (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language 45*, 225–258. 13, 14

van Gompel, R. P. G. and M. J. Pickering (2009). Syntactic parsing. In G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics*, Chapter 17, pp. 289–308. Oxford University Press. 12, 13, 14, 17, 20, 25

van Kesteren, R., T. Dijkstra, and K. de Smedt (2012). Markedness effects in norwegian - english bilinguals: Task-dependent use of language-specific letters and bigrams. *The quarterly journal of experimental psychology 65*(11), 2129–2154. 27, 28, 29

Warren, P. (2013). *Introducing Psycholinguistics*. Cambridge University Press. 24, 37

Wei, L. (2008). Research perspectives on bilingualism and multilingualism. In L. Wei and M. G. Moyer (Eds.), *Research Methodsin Bilingualism and Multilingualism*, Chapter 1, pp. 3–17. Blackwell Publishing. 3, 12, 18

Weinreich, U. (1953). *Languages in Contact: Findings and Problems*. New York, the linguistic circle of New York. 11, 19

Wray, A. and A. Bloomer (2012). *Projects in Linguistics and Language Studies - A Practical Guide to Researching Language,* (3 ed.). Routledge. 46, 48, 50, 51, 71