Eberhard Karls Universität Tübingen
Seminar für Sprachwissenschaft

Bachelor Thesis in Cognitive Science

Reading Noun-Noun Compounds

Early Influences of Compound Frequency and Semantic Transparency

Jonas David Nick

September 3, 2013

Supervisor & First Reviewer

Prof. Dr. R. Harald Baayen
Quantitative Linguistics
Seminar für Sprachwissenschaft
University of Tübingen

Second Reviewer

Prof. Dr. Martin V. Butz
Cognitive Modelling
Department of Computer Science
University of Tübingen

Jonas David Nick
Reading Noun-Noun Compounds
Bachelor Thesis in Cognitive Science
Eberhard Karls Universität Tübingen
Period: April - September 2013

Abstract

This thesis evaluates psycholinguistic theories about the cognitive processing of words. Consequently, the time-course of compound reading is analyzed using generalized additive models in a dataset of eye movements. The theories to be contrasted are sublexical (Taft and Forster, 1975), supralexical (Giraudo and Grainger, 2001) vs. dual route processing (Schreuder and Baayen, 1995) and form-then-meaning (e.g. Rastle and Davis, 2008) vs. form-and-meaning (e.g. Feldman et al., 2009) processing.

As the goal is to find the best model given various predictors, some general mechanisms of eye movements will be demonstrated, e.g. the position in the line has substantial effects, single fixations last longer, are on shorter words, more in the center of the word and influenced differently by frequency measures.

Inspired by Kuperman et al. (2009) it is shown that already the early eye fixations on words are guided by first constituent and compound frequency, providing evidence for parallel dual route models.

Similar to Baayen et al. (2013), Latent Semantic Analysis (LSA) similarity scores (Landauer and Dumais, 1997) permit investigating the time point of semantic processing. The effect of LSA similarity not only shows up in the earliest word fixations, but the data reveals that semantics plays a role even before a word is fixated. In particular, the fixation position in the word is more to the right, when the semantic transparency, i.e. the similarity between compound and second constituent is high. This evidence of parafoveal semantic processing challenges opposing findings obtained with the eye-contingent boundary paradigm (Rayner et al., 1986). In the framework of naive discriminative learning (Baayen et al., 2011), the effect of transparency on fixation position reflects optimization of the landing position for accessing the orthographic information that is most discriminative for the compound.

Keywords: reading, eye-movements, compounds, semantic similarity, morphological processing, generalized additive model

Acknowledgements

First, I would like to thank my supervisor Harald Baayen, whose course on statistical approaches in linguistics drove me to his chair in the first place. He encouraged me to choose a research direction of my interest and continued guiding my thesis research.

His team, the quantitative linguistics group, took me warmly under their wing and had an open ear for all of my questions. I would like to specifically point out the help of Peter Hendrix who conducted the experiments and aided in statistical modelling. I would further like to thank Cyrus Shaoul and Petar Millin who recommended related literature, suggested computational methodologies and gave me further insight to the field of linguistics.

I am very grateful to Martin Butz, my second reviewer, for thinking his way into my research and his fruitful effort to connect related scientific fields. Further, his advices and suggestions at the defence shaped the final outcome of this thesis.

The giants who provided a shoulder to stand on, namely the researchers mentioned in the references earn an extra thanks due to mostly free access to their work. Corpus creators and R package maintainers are specifically responsible for the outcome of this thesis for which I express sincere gratitude. Finally, I would like to thank Xie (2011) for the knitr package that makes the use of R in LaTeXa pure joy.

Contents

1 Introduction

One of the defining properties of language is the possibility to produce and comprehend an unlimited number of different sentences. Since human existence is uncontroversially finite, language makes use of a finite set of primitives manipulated with a finite set of rules. This was arguably first stated by German scientist Alexander von Humboldt who claimed that language makes infinite use of finite means (von Humboldt, 1836, p. 106). An analogy can be made to the set of natural numbers: there is a fixed set of single digits, which can be combined to numbers of unbound length. The unboundedness of language also implies that its possible to come up with a unique sentence. It is rather easy to concatenate words to a valid meaningful sentence until you can be relatively sure that the sentence has never been uttered before.

This design pattern of language poses interesting questions about the exact cognitive basis of language, for which researchers have developed many rivaling models in the past decades. Specifically, those questions range among others from the origin, acquisition, processing, to the production of language. Coming up with answers is inevitable for developing powerful machines, that are capable of imitating the presently unique human ability of processing natural language. Advances find their application in human-machine interaction, information extraction, translation and many others.

Interestingly, these astounding properties of language are not only restricted to the framework of sentences, but also apply to other layers of linguistic analysis, such as the formation of words - the main concern of this thesis. The creation of very long words is possible in many languages. For instance, the word *White house travel office staff* is valid, meaningful and can be extended further, restricted only by your imagination. Some languages have especially long words, such as German where politicians are known for creating word monsters like *Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz* - literally cattle marking and beef labeling supervision duties delegation law. The study of the structure of words provides a window to the inner workings of language processing and mechanisms of the mind.

This thesis will evaluate some of the proposed models of cognitive mechanisms underlying the processing of words with the help of experimental data gathered in an eye tracking experiment, in which the course of reading English sentences was monitored. Therefore, the first chapter recapitulates textbook explanations about the linguistic analysis of words, followed by a chapter concerned about the general observations in the process of reading. If not explicitly referenced otherwise, the information in chapter 1 is taken from Booij (2005) and in chapter 2 from the review article on eye movements in reading by Rayner (1998). Afterwards, the statistical methods used in the analysis of the data are introduced. The next chapter deals with the detailed experimental setup and data processing, leading to the analysis of the experimental data and finally to a general discussion of the results.

2 Morphology

2.1 Terminology

Morphology is the study of word grammar, it deals with the structure of words and its relation to meaning. It is to be distinguished from the linguistic subfields that deal with the intrinsic meaning of words (semantics), the construction of sentences (syntax) and contribution of context to meaning when multiple individuals interact (pragmatics).

The first step of scientific analysis is the categorization of phenomena we observe, hence we should agree on a set of new terms to describe words. This begins with the question of whether two words such as *work* and *works* are actually the same word. On the one hand, both words feel to be instances of the same word. On the other hand, both words differ in their concrete forms. We say that *work* and *works* can be conceived as different **word forms** of the same **lexeme** *work*.

Dictionary makers assume that you have an understanding of language, so they don't list regular word forms such as *works*, because a speaker of English should be able to do that by himself. Instead, they only list the **lemma** *work*, which is the concrete form for the abstract notion of the lexeme *work*.

The building blocks of words are called **morphemes**. In other words, they are the minimal linguistic units with lexical or grammatical meaning. For instance, *work*, -*s* and -*er* are all morphemes. **Simplex words** like *work* are composed of only one morpheme, whereas complex words like *works* are composed of multiple morphemes.

The two important processes in morphology are **inflection** and **word formation**. The difference is that in inflection the lexeme stays the same while word formation leads to a new lexeme. We have already seen an example for inflection when we appended -*s* to the lexeme *work* to gain the third person form. Word Formation, on the contrary, is adding -*er* to the lexeme to receive the noun *worker*, which has a different meaning.

Word formation itself comprises two processes called **derivation** and **compounding**. The difference between the two is that in compounding the building blocks or **constituents** of a word are themselves lexemes. For instance -er is not lexeme, which is why *worker* is an example for a derived word. In contrast, *workbench* is a compound because both constituents *work* and *bench* are lexemes (Figure 2.1).

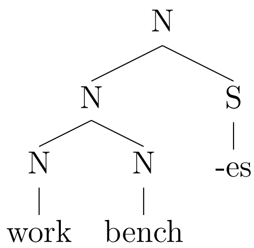

Figure 2.1: The tree depicts the morphological structure of the noun (N) word form *workbenches*. Concatenating the lemmas *work* and *bench* yields the compound lemma *workbench*. Further, the lemma is inflected using the suffix (S) -es.

Since *work*, *works*, *worker* and *workbench* all have a lexeme in common, we call this set the **word family** of *work*.

It would be an impossible task to cover all possible compounds in a dictionary. Rather, it has to be assumed that a speaker of the language is aware of systematicities and thereby able to understand a compound, if it has a transparent meaning. My dictionary lists *bottleneck*, but not *bottle factory*, and indeed everyone would understand *bottle factory* as a factory for bottles. Intuitively, we are able to apply the **compositionality principle** (attributed to Gottlob Frege in Frege (1892)), which states that

(1) 'The meaning of a complex expression is a compositional function of the meaning of its constituents, and the way they are combined.'

On the contrary, *bottleneck* additionally has unpredictable conventionalized meanings.

To exemplify underlying systematicities compare the compounds *soup meat* and *meat soup*. Both have a transparent meaning, one is meat used in soup, the other is soup consisting of meat. We observe here that a compound AB (where A and B are the constituents respectively) denotes a B that has something to do with A. Thus, the constituent A is called **head** and B is called **modifier**. The set of compounds with a single head which contributes most of the meaning are called **endocentric**.

However, the functional relationship between head and modifier can be quite opaque. A *wind mill* is a mill powered by wind, whereas a *flour mill* is a mill that grinds flour. The exact meaning is only accessible in analogy, with contextual information or world knowledge.

In addition to providing the primary semantic information (meaning) the head determines the syntactic category of the compound - i.e. *ice-cold* (Noun-Adjective) is an adjective and *overrate* (PrepositionVerb) is a verb. The head can fulfill additional roles such as determining the gender of the compound in German. For example, the constituents *Fußball* (*football*, male) and *Stadion* (*stadium*, neuter) result in the compound *Fußballstadion* (*football stadium*, neuter).

One may have noticed in previous examples that the head was always at the same position within word. This observation tempted Williams (1981) to propose a rule for the identification of head constituents, known as the **Right-hand Head** Rule:

(2) 'In morphology, we define the head of a morphologically complex word to be the right-hand member of that word.'

In general, most compounds in Germanic languages (English, German, Dutch, ...) obey the rule, but there are a lot of exceptions.

So called **exocentric compounds** completely lack a head so that their meaning can not be easily guessed from their constituent parts. For example, a *pickpocket* is not a kind of pocket but a person who occasionally picks pockets. Similarly, the German word *Rotschopf* (lit. red-mop) doesn't denote colorful hair but the person this hair belongs to. Another category of compounds are **copulative compounds** which lack a single semantic head and whose meaning arises from the relation between constituents that reside on the same hierarchical level. Examples comprise *washer-dryer*, *singer-songwriter* or *bittersweet*.

Until now we have relied on our intuitive understanding of words, but in some cases it can be difficult to distinguish words and phrases. Some languages do not have a written tradition, others do not use the convention that spaces indicate word boundaries. Seprating spaces are merely a matter of **orthography** - how something is actually written - and do not isolate words. And the difference between phrases and words does matter: if you interpret *criminal lawyer* as a noun phrase with an adjective instead of a compound, it denotes a lawyer, who breaks the law, instead of someone who specializes in criminal law.

To be classified as a word, first, it should fulfill a **labelling function**, thereby precisely refering to one specific concept. Second, the criterion of **lexical integrity** (Anderson, 1992) must apply, which says that

(3) 'The syntax neither manipulates nor has access to the internal form of words'.

Therefore, the constituents should always appear in the same order in the compound and it shouldn't be possible to seperate them in a sentence. Further, rules of inflection should not apply to individual constituents. The French *pomme de terre* (lit. apple from the earth - *potato*) clearly has a labelling function, but under the principle of lexical integrity it can not count as a compound, because its plural is *pommes de terre*, whereas the regular plural form of a French word is expressed by appending a suffix at its right edge. A third criterion of word demarcation are language-typical emphasis given to certain syllables in a word. A *dárk room* is a specific place for processing photographic material, opposed to the general notion of a *dark róom*. Fourth, we need context in order to tell compounds apart from noun phrases with attributive adjectives - it is possible to denote lawyers who break the law as *criminal lawyers*.

In this section we tried to define what a word is and how words can be categorized, independently of the language that serves as an example. However, some rules are tendential and only apply to certain languages like the Righthand Head Rule, while others seem to be universal in language like the principle of compositionality. Furthermore, we have seen the complex relationship between a word, its structure and the thing it refers to, leading us to the question for the cognitive basis of these phenomena.

2.2 Morphology and Mind

When it comes to language, the mind is usually thought of as having devices for storing the units of the language and for combining them with the help of rules. On the morphological layer the former is the inventory of words of a language, the **mental lexicon**.

It has several differences compared to a dictionary. First, a mental lexicon is continuously developing, with new words being coined. Further, the mental lexicon also stores the number of occurrence of the word in language, the **word frequency**. This is concluded from the fact that high frequency words are more easily recognized than low frequency words. It can be shown in a **lexical decision** experiment, a common experiment setup in psycholinguistics. A word is presented to the participant and he has to decide whether the word is in the language, while the response latency between the presentation of the stimulus and the push of the button is measured. A typical decision about whether a bi- or triconstituent compound is a known word in the language takes about 763 msec for existing words and 801 msec for nonce compounds (Kuperman et al., 2009). Presenting high frequency words yields substantially faster responses than low frequency words, which is referred to as the **frequency effect**.

In order to measure the frequency of a word, scientists collect large and structured sets of text (**corpora**, sg. corpus) and count occurence of the word. The Web1T (Brants and Franz, 2006) corpus for example collected texts from a trillion web pages and used it to compile frequency lists. In addition to frequencies other databases manually attach information to the corpus. The CELEX database (Baayen et al., 1993) for example contains the morphological structure of the words in the collection. It was critized for frequency measures because it is based only on 16 million words. Instead Brysbaert and New (2009) propose using subtitles of movies, from which they composed a 51 million word corpus.

Another important difference between a dictionary and a mental lexicon is that the only relationship between words in a dictionary is alphabetical listing. On the contrary, words with similar meaning are related in the mental lexicon, which can also be shown in lexical decision experiments. In those setups, first a **prime** word is presented for a short duration of time, followed by a mask, and then the target word, which has to be classified as being in the language or not. If the prime and target word are for example semantically related - such as bread and butter, as opposed to bread and doctor - there will be significantly faster response latencies. This shows that words receive **activation** when accessed in the lexicon, which spreads to related words and enhances their access time.

From the previous section it should be clear that not every complex word needs an entry in the mental lexicon, because of underlying systematicities. In psycholinguistics it is generally a concern what aspects of language are retrieved via storage versus computation. The fact that we can understand and coin new compounds shows that their meaning has to be computed. It is not necessary to store a compound like *bottle factory*, because we can decompose its structure and look up the meaning of individual parts, whereas lexemes with conventionalized and opaque meanings would need to be stored. But there is proof that some high frequency word forms are stored in the lexicon, even if they appear completely regularly. It was shown that the word form shows frequency effects while the lexeme frequency remains constant (Baayen et al., 1997). Thus, recognizing a morphologically complex word can principally take place the following ways:

- **sublexical**: Decomposition into morphemes before whole-word reporesentation is accessed in the lexicon (Taft and Forster, 1975).
- **supralexical**: Morphemes are accessed only after whole word has been accessed in the lexicon (Giraudo and Grainger, 2001).
- **dual route**: Decomposition and whole-word access occur in parallel (Schreuder and Baayen, 1995).
- **connectionist approach**: Morphological effects arise as epiphenomena in mappings between spelling, sound and meaning (Seidenberg and McClelland, 1989; Seidenberg and Gonnerman, 2000).

We are not going to discuss connectionist models further because the other approaches are far more popular in morphological processing. Sublexical and supralexical refer to the point in time when morphological processing takes place - either before (sub-) or after (supra-) lexical access. Evidence of sublexical models was for example given by Forster and Taft (1976) who compared lexical decision latencies of compound nonwords whose first constituent is a word (e.g. *dustworth*, *footmilge*) with compound nonwords whose first constituent is not a word (e.g. *trowbreak*, *mowdflisk*). The former took significantly longer to reject, suggesting that the compound nonwords are decomposed, because otherwise both conditions should evoke the same latencies.

Supralexical processing was shown for example by Giraudo and Grainger (2001) who conducted a priming experiment. They stated that according to the sublexical hypothesis the lexical decision latency should take longer for target words (e.g. *taxation*) to identify if the prime word is a derived word (*taxability*) than the root (*tax*) because the former prime requires extra parsing. Their results indicate that this is not the case and so they argue for a supralexical account.

The Experiments conducted for this thesis will provide evidence that strict supralexical or sublexical models are questionable and of the previously suggested models we should favor a dual route model.

The second topic of this thesis is whether morphological processes take semantics into account. The **form-then-meaning** (e.g. Rastle and Davis, 2008) view states that complex words undergo a two-step process: initially an early meaning-blind decomposition followed by semantic interpretation. One source of evidence is that in priming studies Rastle et al. (2004) found no difference between conditions where the prime and target word were morphologically related (e.g. *trucker* - *truck*) and conditions where prime and target only seem to be morphologically related (e.g. *corner* - *corn*). A second source of evidence for form-then-meaning processing comes from neuroimaging data that was gathered in primed or unprimed lexical decision tasks. Lavric et al. (2007) for instance found that semantically transparent primes (*truck* - *trucker*) show a similar ERP signal to semantically opaque primes (*corn* - *corner*) at 340 to 500 msec after stimulus onset, which is what they count as evidence that morphological decomposition is independent of semantics. In addition, they revealed that the signal between 220 and 260 msec differed significantly between transparent/opaque and only orthographically related primes. Taken together with other brain imaging studies, semantic interpretations seem to occur between 300-500 msec after stimulus presentation and the signature of morphological processing is found substantially earlier.

However, several studies with a different experiment setup, namely with eye tracking, are incompatible with the late locus of semantic effects. For isolated words and words in context lexical and semantic factors affect the earliest eye-movements in the first 200-250 msec after stimulus presentation. Lexical frequency of the compound as well as the family size of the left constituent affect the duration of compound processing (Kuperman et al., 2009). Also, Marelli and Luzzatti (2012) showed that semantic transparency has a reliable influence on early eye-movements.

The experiments conducted for this thesis make likewise use of eye tracking to analyze the time-course of compound processing in order to shed light on the timing of semantic effects. In line with Baayen et al. (2013) it will be argued

Table 2.1: LSA similarities from `http://lsa.colorado.edu/` using the General_Reading_up_to_1st_year_college (300 factors) space and term to term comparison.

Text	Text	Similarity
garbage	trash	0.86
corner	corn	0.1
dark	darkness	0.78
honeymoon	moon	0.01
half-moon	moon	0.96
network	work	0.08
baseball	ball	0.69

for the **form-and-meaning** hypotheses and will go further by showing that semantic effects occur even before the whole compound has been scanned. Before we proceed with familiarizing ourselves with the general mechanisms of eye-movements in the next chapter, it will be clarified in the next section how we operationalize the semantic effects that we try to find.

2.3 Semantic Similarity between Compounds

A common way to measure semantic similarity between words is by using **Latent Semantic Analysis (LSA)** (Landauer and Dumais, 1997), a computational technique that estimates similarity based on the contexts in which the words appear by using large corpora. To calculate the similarity with LSA, first, the texts are represented as a matrix where every row stands for a unique word and every column for the place it occurs. Latent Semantic Analysis assumes that terms close in meaning will occur in similar pieces in text. Second, singular value decomposition is applied to the matrix, which boils the matrix down to the relevant columns. Then, the similarity of two words is computed by measuring the cosine of the angle between two vectors formed by the rows of the words. The similarity scores range from -1 to 1, a higher score stands for a shorter distance between the vectors and represents a higher similarity between the words.

To convince yourself of the authors' claim and many others that the similarity score approximates human judgements, you can study exemplary similarities in table 2.1. The difference in similarity between *honeymoon/moon* and *half-moon/moon* is of particular interest. It shows that we can treat the scores between the whole compound and the right constituent, which most often serves as the head, as an estimation of semantic opacity. Baayen et al. (2013) compared the similarity ratings between constituent and compound computed by LSA with scores they obtained from human raters. They showed that the Pearson's correlation yields $r = 0.51$ for left-whole LSA similarity and $r = 0.44$ for right-whole LSA similarity. Further, they showed that left-whole and right-whole LSA similarity is more predictive for lexical decision latencies than human judgements.

3 Reading

Reading feels like a continuous flow along the sentence. Actually, when measuring the place where the eyes are looking to with an **eye tracker**, it becomes evident that reading involves periods of rest and sudden movement. An eye tracker typically measures eye movements with a video camera, spots the pupil center and computes the point where the participant is looking. The device reveals that the eye moves rapidly between **fixation positions** where the eye remains still for around 200 ms (Figure 3.1) Because the eyes' sensitivity to

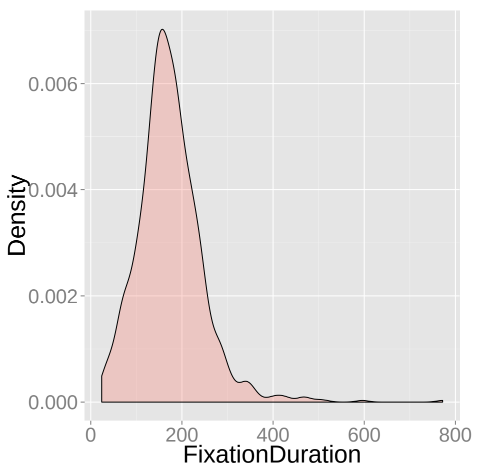

Figure 3.1: Density of fixation duration in our eye tracking dataset.

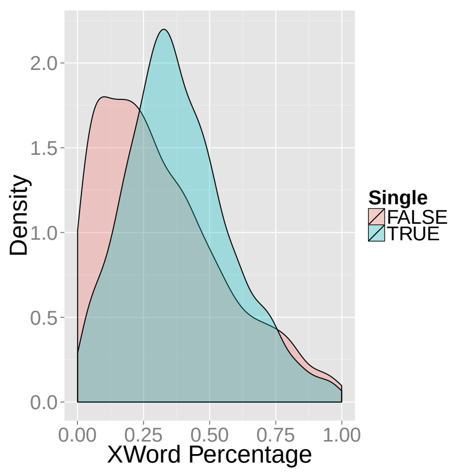

Figure 3.2: Density of first fixation positions by single fixations.

light is maximal in the **fovea** and decreases in **parafoveal** and peripheral regions, the acuity (clearness of vision) is only optimal around fixation positions. Interestingly, we find a frequency effect in reading, high frequency words take up less time . This shows that properties of the word influence fixation time, such that fixations can tell us about cognitive processing. The eye movements are referred to as **saccades**, they are on average 7-9 letters long, range from 1 to over 15 letters, and last typically 30 ms. During saccades no visual information is taken up, a phenomenon called saccadic suppression. Figure 3.3 depicts a sequence of fixation positions in a sentence. 15% of content words and 65% of function words are not fixated once during reading, indicating that parafoveal vision and expectations about the sentence structure are sufficient

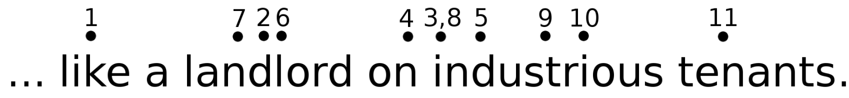

Figure 3.3: Section of the Brown corpus with fixation positions of a participant.

to identify the word. Some words only receive a **single fixation**, while others are fixated multiple times. We define **first fixations** as both single fixations and the first-of-many fixations in a word, as opposed to **refixations** which happen mostly with words we find complicated. In figure 3.2 we can see the distribution of first fixation positions and that the first fixation position in a word differs between single and non single fixations. Moreover, we can see that most first fixations are in the first half of the word. In addition to a continuous reading flow from left to right, about 10-15% of saccades are **regressions**, eye movements which go back in text. Short regressions suggest that the reader has problems processing the word, while longer regressions imply that the reader did not understand the text.

Cognitive processes accompanying reading and operationalized by fixation time involve many subprocesses including saccade programming, word processing and integration into the sentence. It takes at least 150-170 ms to plan a saccade, which suggests that saccade programming is done in parallel with comprehension processes. The classic model of eye movement control was proposed by Morrison (1984). It states that the word in the fixation, let's call it n, is currently processed. After completion of word processing, attention shifts to word $n + 1$ and initiates to prepare a motor program that brings $n + 1$ in foveal view. Meanwhile $n + 1$ is processed parafoveally until finally a saccade to $n + 1$ is executed. If $n + 1$ is already identified before it is fixated, attention shifts to the word $n + 2$ such that $n + 1$ will be skipped. From this perspective fixation times reflect lexical access, whereas regressions occur because postaccess processes like semantic and syntactic integration intervene. A word has to be fixated again if the acuity is not high enough to identify the letters at the end of the word, which explains the higher probability of refixations on longer words.

To investigate the amount of parafoveal information uptake researchers have been using the eye-contingent boundary paradigm, in which participants where presented with a sentence, where the target word is replaced when a fixation lands on the word. Thus, the gaze duration on the target word provides insight into how much the word has already been processed. Rayner et al. (1986) conducted a study, where the target word is orthographically related (*sorp*), semantically related (*tune*) or unrelated (*door*) before it is fixated and replaced by the correct word (*song*). It was shown that the orthographic condition causes similar gaze durations on the target word as the control condition, whereas the semantic condition resembles the unrelated condition. Therefore, the view on parafoveal information uptake has been that only first letters and the length of $n + 1$ are received (Bertram, 2011). However, recent research has shown that word processing is not constrained to n, but upcoming words $n + 1$ and $n + 2$ are processed parallel even though they are only projected to parafoveal regions in the eye (Kliegl et al., 2007). More importantly, Hohenstein et al. (2010) demonstrated that semantically related parafoveal words can give a preview benefit on the current fixation which challenges previous models.

All in all, the first fixation duration is able to tell us something about lexical processing, and thus about morphological processing, if the fixation lands on a morphologically complex word. Moreover, by examining the first fixation position we can reveal properties of parafoveal processing, because planning of a saccade can not be informed by previous foveal processing. In contrast to many of the experiments with a factorial design that were introduced, here we will conduct a regression study which will be explained in the next chapter in more detail.

4 Statistics

In order to infer the relevant features of words from the eye tracking data, we will use statistical modelling, namely regression analysis. The goal of regression analysis is trying to estimate the relationship between predictor variables and a response variable. The standard model, simple linear regression, assumes that the response variable y can be represented as a weighted sum of predictors x_i:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$$

The variables β are called **parameters** of the model and β_0 is the **intercept**. The parameters are normally estimated by minimizing the squared error between the model predictions and the actual values.

The advantage of a regression design in contrast to a factorial experimental design is that the predictor variables do not have to be artificially dichotomized. Rather, numerical variables appear in their full spectrum in the model, which leads to increased statistical power (Cohen, 1983; Baayen, 2010).

Generalized additive models (GAMs, Hastie and Tibshirani, 1986) relax the assumption of linearity in the partial effect of predictor variables. In addition to multiplying the predictor, they allow rather flexible specification in terms of **smooth functions** f_j, e.g.

$$y = \beta_0 + \beta_1 x_1 + f_1(x_2) + f_2(x_3, x_4) + ...$$

The smooth functions f_j are represented as regression splines, i.e. a sum of basis functions. In order to prevent too complex functions and overfitting of the data, the degree of smoothness is controlled by penalizing the addition of basis functions so that cross validation (training on subsets of the data) shows good fits. Figure 5.1 for example depicts a smooth function with the predictor on the x-axis and its partial effect on the y-axis. In addition to main effects the GAM is able to take non-linear interactions between multiple variables into account (c.f. f_2) by using **tensor products**. The graphs for these interactions (e.g. 5.2) show the fitted values as a function of the predictors participating in the interaction on the x- and y-axis. The color code depicts the fitted values, where green is lowest and pink/white is highest. We can fit all parameters of the model by factorial predictors, so that we have estimates for each factor level - for instance having one curve for single fixations and one for non-singles.

GAMs allow for **random effects**, in which there is a different intercept for every factor level of the predictor. This is useful when the set of possible levels for the predictor is not fixed and each of the levels is non-repeatable. In contrast to predictors such as drug/no-drug or isSpaceSeparatedWord, the word items and the subjects of an experiment items are best represented by random intercepts.

GAMs provides probabilities whether an effect is present, i.e. its estimated parameter differs from zero. Specifically, the obtained **p-value** is the probability of estimating this parameter given that it is actually not present in the data. In addition to testing for significance of the predictors (p-value < 0.05), we assess the quality of the model with the **Akaike Information Criterion** (AIC). The AIC gives an estimate of the information lost when the given model is used to represent the process that generates the data in comparison to the (unknown) true process. Therefore, the lower its value, the better the model. Moreover, it penalizes the usage of many predictors, which helps to prevent overfitting.

In the beginning of the modelling process, random forests were used to gather an overview of variable interactions and estimate variable importance. All of the following models were implemented in the statistical computing language R (R Development Core Team, 2011) using the mgcv package (Wood, 2006) for GAMs and the party package (Strobl et al., 2008) for random forests.

5 Experiment

The main concern of this work is, first, find the best model given our predictors to learn about general mechanisms in reading.

Second, we try to replicate the findings of Kuperman et al. (2009), who showed that compound and first constituent frequency as well as family size governs early fixation duration in reading, which makes strictly sub- or supralexical models of morphological processing unlikely. In their set up the participants pursued a lexical decision task while wearing an eye tracker. In contrast to our dataset the words were presented in isolation, thus, if we are able to confirm their results it would increase the ecological validity.

Third, Baayen et al. (2013) showed that LSA similarity between constituents and compound influences first fixation duration and first fixation position, which dismisses staged models and argues for form-and-meaning morphological processing. Their findings indicate that words already undergo a semantic analysis in parafoveal vision, which is so controversial that more data and potentially evidence should be gathered.

The two sources can be compiled to the following hypotheses:

(1) Both immediate constituents and the whole compound affect lexical processing of compound words.

(2) Compound Frequency, left constituent frequency and left constituent family size show up at the first fixation and precede right constituent frequency and family size effect.

(3) LSA similarity predicts the durations of first fixations.

(4) LSA similarity predicts the position of first fixations.

To investigate the hypotheses, a dataset of fixations collected with an eye tracker is used. In order to limit difficulties when trying to detect and analyze different compound types, we restrict ourselves to noun-noun compounds. Then we try to find the best model fit using GAMs for fixation duration and fixation position given the predictors, so that the significance and estimates of predictors cast light on the hypotheses.

5.1 Data Collection

The fixation duration and fixation position data were gathered in an eye tracking experiment conducted by Peter Hendrix. With an eye tracker from the EyeLink1000 series with chin-rest and a sampling rate of 500 Hz the fixation durations and positions of the right eye were measured, while participants read the entire fiction section of the Brown Corpus (Francis and Kucera, 1979), which is in English. This section contains contains 126 text consisting of 35-50 pages each. The monitor displayed full pages, which are up to 10 lines long. The text appeared on a monitor with 1024 x 768 resolution and was presented in a black Courier New font with a size of 24 pixel on top of a white background. Each page was preceded by a fixation mark at the top left of the screen, at the same location as the first letter of the first word of the text, to correct for minor movements of the head.

There were 4 participants, graduate students at the Linguistics department of the University of Alberta, consisting of two males, aged 26 and 31, and two females, aged 28 and 26. Three participants were native speakers of English, one was near-native (the 26 year old female). They received \$20 per hour for their participation and a \$250 bonus for completing the entire experiment which took around 100 hours. A game pad allowed moving to the next page and self-calibration every 5 pages with 9 reference points. Participants read 2 texts per hour with 5 minute breaks between the texts. If they decided to run themselves for more than one hour at a time, they took 10 minute breaks between each session of 2 texts. They were instructed to move as little as possible and read at a natural pace.

A post-processing step was necessary because sometimes the vertical positions of fixations were inaccurate, mostly at/near the ends of lines. Fixations were therefore corrected by a correction algorithm, which compared the x and y coordinates of preceding and following fixation. The algorithm put about 98% of fixations on the correct line, such that its output was inspected by student assistants and manually corrected where necessary.

In total, the dataset consists of 920.000 fixations on 253.000 words.

5.2 Data preparation

In order to extract only fixations on noun-noun compounds from our dataset, every word that occurred in the text was looked up in the CELEX database. CELEX allows us to, first, map the word form to a matching lemma and then obtain the morphological structure of the lemma. The structure enables obtaining the individual constituents that make up the compound. If there were multiple lemmas for a word form, every one of them was checked whether its a noun-noun compound. Compounds which are separated by spaces needed special treatment, such that every word bigram (two adjacent words) without intervening delimiters in the text was looked up in the database.

There are obvious problems with this naive approach. First, if words have the same spelling but different morphological structures, one of which is a noun-noun compound, then the word is marked as a noun-noun compound even if sentence structure or context would reject the compound interpretation. For example, *foxhunt* is listed in CELEX both as a noun-noun compound as well as a noun-verb compound. Second, two adjacent words could be considered a compound according to a CELEX, even if they do not represent a compound. The noun-noun detection algorithm returns a false positive for the sentence *'He was parking lots of cars.'*, because it finds *parking lot* as a word form of a noun-noun compound. However, by manual inspection it was ensured that those problematic cases are rare and are therefore only slightly able to bias the quantitative analysis. This does not rule out the possibility of false negatives at all, because we can only detect compounds that were listed in CELEX, i.e. it appeared more than once in their corpus. This is problematic given the vast number of possible noun-noun compounds.

In order to fit reliable models of natural data it is often necessary to remove data points, which can be considered outliers, because they are numerically distant from the rest of the data and thus not governed by variables we want to investigate but by measurement errors or participants short-term inability to stay with the experiment. It was derived from theory that fixations which last less 70 msec are to be considered misfixations and should not be analyzed further, because this time frame is too small to go beyond simple visual uptake. 237 measurements were taken out of the dataset due to this train of thought. On the other side, there were 18 fixations which were erased because they seemed way too long for normal fixations because they lasted longer than 520 msec (more than three standard deviations away from the mean) and distorted the quantile-quantile-plot.

The subset of noun-noun resulted in 5475 fixations on 944 unique words. Of all those fixations, there were 2002 single fixations, 1505 first of many fixations and 1968 last of many fixations.

5.3 Predictors and Response Variables

We have some predictors that stem from the presentation, namely Line, X and XParagraph. The text for the participant is presented in paragraphs, which span multiple lines and may consist of multiple sentences. The numerical variable **Line** denotes the line number of the current fixation. **X** is how far the fixation is into the line and **XParagraph** shows how much the fixation is into the line if the whole paragraph was in one line. The variable **InSentence** gives the position in the current sentence in percent.

The following predictors are dependent on properties of the word. **Length** stands for the length of the compound in number of letters. **CompoundFrequency** denotes the lemma frequency of the compound in the Web1T corpus. **Const1Frequency** and **Const2Frequency** stand for the frequency of the first and second constituent in the subtitle corpus. Originally, all frequency measures were intended to be from the CELEX corpus, but the subtitle corpus has been proven as making more accurate predictions for lexical decision latencies. In addition, CELEX does not list frequencies of space-separated lemmas, due to limits of computational power at that time. The subtitle corpus also does not have frequency list for those compounds, which is why frequency lists of Web1T were used.

It has been shown that the family size, i.e. the number of lemmas that share one constituent with the compound, influences fixation duration (Kuperman et al., 2009). Naturally, family size is highly correlated with the frequency of the constituent. Therefore, two linear models were fitted where the family size of the respective constituent was predicted given the constituents frequency. The residuals of both models yielded the two predictors **Const1ResidFamSize** and **Const2ResidFamSize**.

In order to evaluate semantic effects on early measurements, the predictors **LSACompConst1**, **LSACompConst2** and **LSAConst1Const2** were introduced, which denote the LSA similarity between compound and constituent and the constituents with each other. The similarity effects were obtained from `http://lsa.colorado.edu` using the General_Reading_up_to_1st_year_college (300 factors) space and term to term comparison. By manual inspection it was realized that LSACompConst1 and LSACompConst2 are not reliable for compounds with a separating space, which is why we always include a by=Space part in the model's LSA term and ignore the effects for space-separated compounds. LSAConst1Const2 remains meaningful nevertheless.

The variable **PreviousFixationDuration** is self-explanatory, the variable **Space** marks compounds that are separated by spaces and **Single** is used to indicate fixations that were the only ones on that particular word. In addition, there can be random intercepts for **Subjects** and **Words**.

Including all predictors, the collinearity index κ computed by the languageR package (Baayen, 2011) is 12.6867. This means that they are not harmfully correlated and do not suppress each other.

The first response variable to be studied is the **fixation duration**. It was log-transformed to reduce the skewness of their distribution and to reduce the influence of atypical outliers as suggested by Baayen and Milin (2010). The raw distribution of the variable is depicted in figure 3.1. The fixation position is the number of pixels the participant fixated horizontally into the word and therefore the variable is coined **XWord**. The related variable XWord percentage (not included in the models) can be inspected in figure 3.2.

Table 5.1 provides the list of variables, as well as their range, mean and standard deviation. In addition to the fixation duration the frequency mea-

Table 5.1: Descriptive statistics on continuous predictors and response variables.

Variable	range	mean	sd
X	43:942	431	238
XParagraph	45:7023	2939	1793
Line	1:8	3.6	1.9
InSentence	0.02:1	0.64	0.29
CompoundFreq	-3.0:2.5	0	1
Const1Freq	-2.5:2.3	0	1
Const2Freq	-2.4:1.7	0	1
Const1ResidFamSize	-2.1:1.9	0	0.6
Const2ResidFamSize	-2.0:2.6	0	0.5
LSACompConst1	-0.07:1	0.4	0.26
LSACompConst2	-0.09:1	0.3	0.28
LSAConst1Const2	-0.03:0.92	0.22	0.15
PreviousFixationDuration	20:784	177	70
FixationDuration	4.2:6.2	5.2	0.3
XWord	0.1:240	68	43.5

sures require a log-transformation, because their original distributions follow a power law, which would make them unsuited for regression analysis. The frequency measures, and the family size residuals were standardized, such that their mean is zero and the standard deviation is one. All other variables denote their original values.

5.4 Analysis

The general process for the analysis of the dataset was, first, include many predictors and interactions of predictors. Then, sequentially dismiss predictors if they are not significant or do not reduce AIC substantially.

5.4.1 Single Fixation Duration

We start testing our hypotheses with the analysis of the duration of single fixations with a generalized additive mixed model (Table 5.2).

First, we can observe that the manner of presentation does matter, specifically Line and X are highly significant, as depicted in the p-value column. The higher the current line number the longer the fixation (Estimate column), probably because the paragraph unfolds its complexity further in the text, which elicits longer fixations. The partial effect of X can be seen on the vertical axis of figure 5.1 for different values of X. Fixations in the beginning of

Table 5.2: Generalized additive mixed model fitted to the single fixation duration, reporting parametric coefficients and estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), F and p values for the tensor products and random effects.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	4.9899	0.1872	26.6600	< 0.0001
Line	0.0158	0.0036	4.3931	< 0.0001
Const1ResidFamSize	0.0260	0.0121	2.1492	0.0317
B. smooth terms	edf	Ref.df	F-value	p-value
Subject	2.7580	3.0000	11.2412	< 0.0001
Word	44.4680	356.0000	0.1608	0.0063
X	5.0855	6.1684	18.0410	< 0.0001
te(CompoundFreq,LSACompConst2):Space	3.8378	4.2296	3.1268	0.0127
te(CompoundFreq,LSACompConst2):no Space	3.0000	3.0000	2.8470	0.0363

the line last longer than in the end of the line. Because neither InSentence nor XParagraph are able to replace the X variable, we can conclude that the longer fixations stem from the confusion when the participants change into a new line.

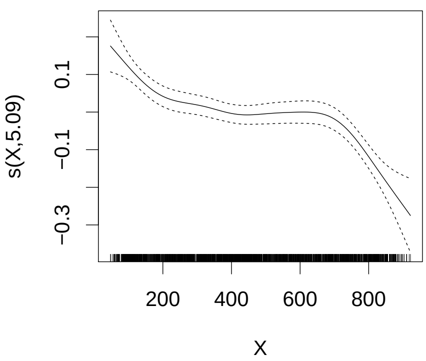

Figure 5.1: Smooth partial effect of X on single fixation durations. The dashed lines depict 95% confidence intervals.

Further, we can see that there are highly significant random intercepts for subject and word, which shows that fixation duration depends on individual differences, such as experience, as well as properties of the word that are not captured by our variables.

As we have stated earlier, it is nothing unusual that we find frequency effects for fixation durations, especially in single fixations, because after the fixation the lexical processing should be done. In fact, CompoundFreq without the interaction is highly significant ($p < 0.0001$) and shortens the fixation duration. The first constituent elicits no frequency effect, but still has an influence on fixation duration, namely a high family size lengthens the fixation duration. This opposes the findings of e.g. Bertram et al. (2000), who found in several experiments that family size actually shortens fixation duration. Despite the significance of family size here, it does not reduce AIC extensively, but only by roughly 3 units. However, the results of Bertram et al. (2000) were obtained in lexical decision experiments, where rather different processes are at work. For example, it would be reasonable to assume that higher family sizes increase uncertainty about the compound which leads to longer fixations. The predictor Const1FamSize, where the confounding influence of the frequency variable has not been removed, does not become significant.

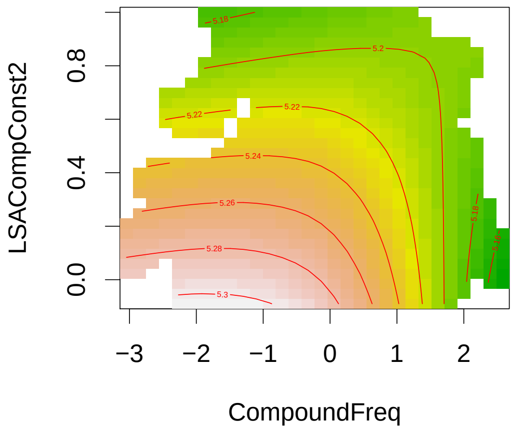

Figure 5.2: Tensor product smooths for the interactions of compound frequency by whole-right LSA similarity for single fixation durations of compounds without spaces. Green areas stand for shorter fixation duration and pink areas for longer fixations.

It seems that semantic similarity has an influence on word processing, because adding LSACompConst2 to the model reduces AIC by $\tilde{1}$5 units and the interaction with CompoundFrequency is significant. In Figure 5.2 we can see that relatively low frequency of compound and low semantic similarity between head and compound lead to longer fixation durations. Further, the interaction functions like an OR gate - if one of the predictors has a high value, the other one doesn't matter. As stated before, the interaction for space-separated compounds is not reported here because LSA similarity is not reliable for those words. In contrast to these results, Baayen et al. (2013) found that the interaction of LSAConst1Const2 by Const2Freq is significant - here we do not find a second constituent frequency effect at all.

5.4.2 First-of-many Fixation Duration

Table 5.3: Generalized additive mixed model fitted to the first-of-many fixation duration.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	5.2102	0.1011	51.5537	< 0.0001
CompoundFreq	-0.0232	0.0099	-2.3441	0.0192
Const1FreqCelex	-0.0223	0.0091	-2.4473	0.0145
B. smooth terms	edf	Ref.df	F-value	p-value
Subject	2.7405	3.0000	10.3629	< 0.0001
X	3.6015	4.5123	2.2912	0.0504
te(XWord,LSACompConst2):no Space	6.0218	6.8550	12.8614	< 0.0001
te(XWord,LSACompConst2):Space	4.8355	5.5046	0.9714	0.4361

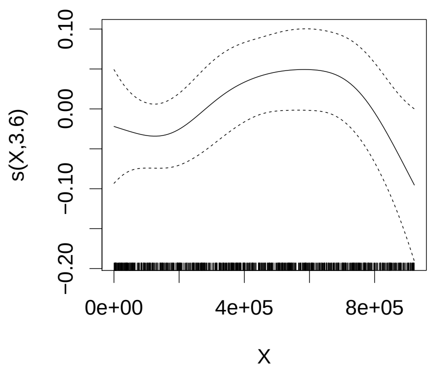

Figure 5.3: Smooth partial effect of X on first-of-many fixation durations.

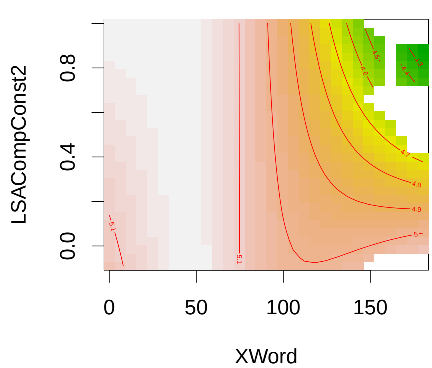

Figure 5.4: Tensor product smooths for the interaction of fixation position in the word by whole-right LSA similarity for first-of-many fixation durations of non-space separated compounds.

Next, we are considering first-of-many fixation durations (Table 5.3). As before, Subject is a random effect and X has a highly significant effect. The partial effect of X shows (Figure 5.3) that fixations at the beginning of a line are especially short, which probably indicates that after jumping from the preceding line, there are many short fixations necessary to focus on the current line. Also, it sticks out that individual Words do not have significant random intercepts as it was in the single fixation case. Also, Line is not significant. Because the first-of-many fixation is an early measurement (otherwise no refixation would be necessary) and the first constituent is projected to the fovea, we expected its frequency to be significant (Hyönä et al., 2004). Only the frequency measurement from CELEX is significant, SUBTLEX frequencies fail to achieve a significant effect ($p \approx 0.12$). However, that the compound frequency reaches significance means that the full compound is recognized even though a refixation is necessary. This goes well with the findings of Kuperman et al. (2009), who claimed that this early full-form access poses a problem for sublexical models. The importance of the compound frequency for predicting first-of-many durations was supported by fitting a random forest to the data, which showed that CompoundFreq was the third most important predictor. Both frequency measures facilitate lexical processing.

Further, the interaction of XWord with LSACompConst2 is significant for compounds without separating spaces. XWord alone is significant, but the interaction with LSACompConst2 reduces AIC by 48 units. Figure 5.4 shows that the farer the first-of-many fixation is into the word, the shorter its duration. Because there are multiple fixations necessary to process the word, this does not mean that lexical processing is facilitated by a further jump. Semantic similarity influences this effect for higher XWord by shortening the fixation duration. Baayen et al. (2013) found a main effect of LSACompConst2 (higher values lead to shorter fixations), whereas in this experiment semantic similarity only occurs in an interaction.

5.4.3 First Fixation Duration

We get all first fixations by taking both the single and first-of-many fixations into account in order to find a model that captures their similarities (Table 5.4).

Table 5.4: Generalized additive mixed model fitted to the first fixation duration.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	4.9866	0.0701	71.1465	< 0.0001
Line	0.0106	0.0027	3.8520	0.0001
PreviousFixationDuration	0.0348	0.0126	2.7633	0.0058
Single	0.0882	0.0111	7.9836	< 0.0001
Const1ResidFamSize: Single	0.0021	0.0128	0.1652	0.8688
Const1ResidFamSize: no Single	0.0288	0.0116	2.4856	0.0130
Const1Freq: Single	-0.0157	0.0089	-1.7559	0.0792
Const1Freq: no Single	-0.0034	0.0079	-0.4318	0.6659
B. smooth terms	edf	Ref.df	F-value	p-value
Subject	2.8771	3.0000	31.5248	< 0.0001
Word	50.0336	398.0000	0.1586	0.0043
X: no Single	8.3836	8.8806	17.2234	< 0.0001
X: Single	7.0571	8.1089	15.6929	< 0.0001
te(CompoundFreq,LSACompConst2):no Space	3.6384	3.9564	4.1148	0.0027
te(CompoundFreq,LSACompConst2): Space	3.4904	3.7935	0.6317	0.6314

First, we note that Subject and Word have random intercepts, Line occurs as a main effect and X as a smooth term. The graphs for X are not surprisingly similar to the graphs we have seen in the individual analysis of single first-of-many fixations. Because we now have more data points at hand, we find more subtle effects like the previous fixation duration, which spreads to the current fixation. Moreover, we see that Single is a main effect, such that single fixations last longer than first-of-many fixations. For single fixations a high family size of the first constituent again seems to make comprehension more difficult.

This model seems to only slightly suggest that the first constituent frequency could play a role and if so, then only for first-of-many fixations. However, if we cut out the interaction CompoundFreq by LSACompConst2 and compare CompoundFreq alone (highly significant) with CompoundFreq in interaction with Const1Freq, the latter reduces the AIC by ≈ 60 units. In the current model the interaction between CompoundFreq and LSACompConst2 is shown, which reduces the AIC also by ≈ 60 units compared to CompoundFreq alone. The significant part of the interaction captures both single and first-of-many fixations and we see in figure 5.5 that the semantic similarity effect is weaker than in the single fixation only model and only mid-to-lower similarity eases reading.

In summary, the model shows that single and first-of-many fixation durations differ substantially in most measurements, the only predictor both fixa-

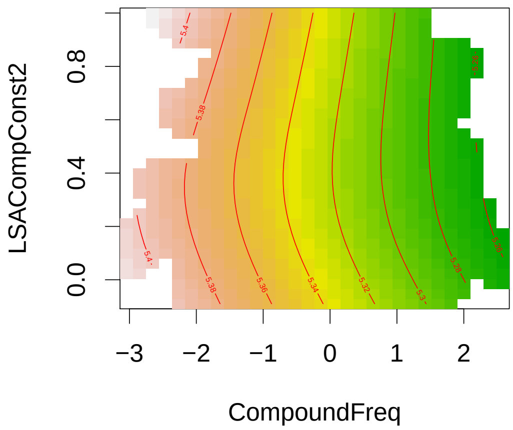

Figure 5.5: Tensor product smooths for the interaction of compound frequency by whole-right LSA similarity for single and first-of-many fixation durations of non-space separated compounds.

tion types have in common is the interaction of CompoundFreq by LSACompConst2.

5.4.4 Probability of Refixation

The next model tries to predict whether an additional fixation on the compound is necessary, predicting the probability of a first-of-many fixation against a single fixation of all first fixations (Table 5.5).

Table 5.5: Generalized additive mixed model fitted to the probability of a refixation.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	1.4531	0.9293	1.5636	0.1179
CompoundFreq	-0.3363	0.0574	-5.8572	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
Subject	2.5740	3.0000	22.2960	< 0.0001
Word	70.0010	400.0000	97.3006	< 0.0001
X	4.2698	5.2380	22.1164	0.0007
InSentence	2.8453	3.4720	10.0504	0.0275
XWord	5.1162	6.2261	169.0033	< 0.0001
te(Length,LSACompConst2):no Space	3.0001	3.0002	29.5085	< 0.0001
te(Length,LSACompConst2):Space	3.4568	3.7833	21.8129	0.0002

We first recognize that Subject and Word have random intercepts like in the preceding models. Further, CompoundFreq appears as a linear main effect for which high values make it less likely to result in refixations.

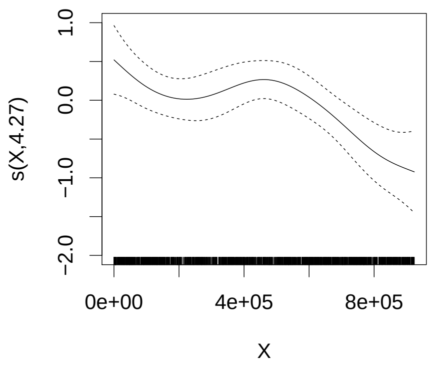

Figure 5.6: Smooth partial effect of X on probability of refixation.

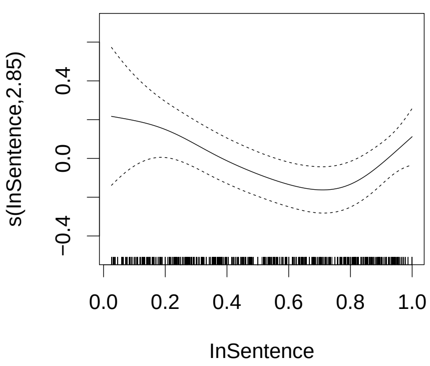

Figure 5.7: Smooth partial effect of InSentence on probability of refixation.

Figure 5.6 shows that readers most likely refixate at the beginning of a line, which goes well with our previous findings about X. We can see in figure 5.7 that readers refixate less often in the middle of a sentence. If we stick to the hypothesis that refixations occur on complicated words or sentence structure, we can conclude that most of the magic happens at the beginning or at the end of a sentence. Figure 5.8 suggests that there is a sweet spot for single fixations at the middle of a word - if it is fixated at the beginning or at the end it will rather result in a refixation. This spot reflects the optimal viewing location which has often been found in eye tracking studies.

Unsurprisingly, longer compounds are more likely to receive multiple fixations (Figure 5.9), but Length is further in interaction with LSACompConst2, which reduces AIC by 100 units. For longer compounds a higher similarity between head and compound leads to a higher probability of a refixation. It is necessary to note, that if we would truly want to explain the moment just before a refixation is made or not, we would have to include the fixation duration as a predictor. Including this predictor, however, would eliminate the effect of LSA similarity, i.e. the influence of LSACompConst2 seems to be explained by FixationDuration. Instead of an effect of LSACompConst2 Baayen et al. (2013) found a three-way interaction of XWord, Const2Freq and LSAConst1Const2. The interaction would also become significant in this model, but it only reduces AIC by 15 units here, which is considered too low for such a complicated term.

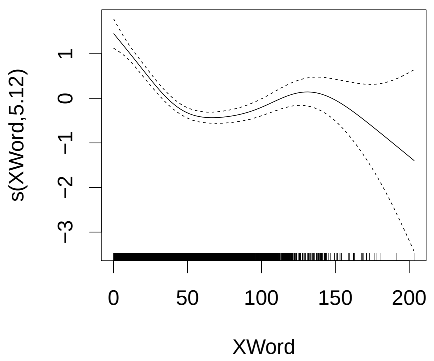

Figure 5.8: Smooth partial effect of XWord on probability of refixation.

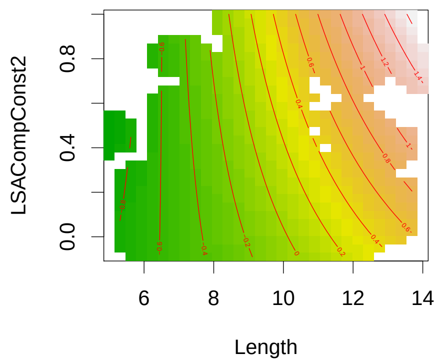

Figure 5.9: Tensor product smooths for the interaction of length by whole-right LSA similarity for single and first-of-many fixation durations of non-space separated compounds.

5.4.5 First Fixation Position

To investigate the parafoveal processing in reading, we continue with investigating the influences on the first fixation positions in the compounds (Table 5.6).

Table 5.6: Generalized additive mixed model fitted to the first fixation position.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	53.9836	14.0791	3.8343	0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
Subject	2.9460	3.0000	54.7067	< 0.0001
Word	80.8632	396.0000	0.2935	< 0.0001
X	8.6936	8.9632	26.0037	< 0.0001
PreviousFixationDuration	3.4603	4.3415	7.5270	< 0.0001
te(Length,LSACompConst2):no Space	3.4206	3.6507	7.8358	< 0.0001
te(Length,LSACompConst2):Space	6.0681	7.1253	1.7551	0.0905
te(Const1ResidFamSize,LSAConst1Const2)	4.5568	5.0181	3.0333	0.0097

It is evident that there are again random intercepts for subject and word, as well as effects of X and PreviousFixationDuration. In figure 5.10 we see that at the beginning and end of a line the participants fixated especially far into the compound. The first peak corresponds to the short fixations at the beginning of a line, because participants have to figure out first how to begin reading the line. Figure 5.11 shows that the higher the previous fixation duration the more cautious the reader and thus he does not fixate far into the word.

Similar to Baayen et al. (2013) we see an interaction between Length and LSACompConst2 in figure 5.12. But in the Baayen et al. (2013) the effect of LSACompConst2 is that a higher similarity leads to earlier fixations, whereas here LSACompConst2 leads to farer fixations for longer compounds. Also, their effect of length appears to go directly into the opposite direction, because they reported that shorter compounds have their fixations more at the beginning of the word. The reason for this becomes visible when we use the percentage position in the word instead of XWord as a predictor. Then, the surface of the interaction shows that longer compounds are relatively more fixated in the left while the effect of LSACompConst2 stays basically the same (Figure 5.14). But the absolute fixation position of longer compounds is more to the right. Despite inclusion of LSACompConst2 reduces AIC by extraordinary 800 units, it is not a main effect as opposed to Baayen et al. (2013).

Moreover, they found an interaction of Const2Freq by LSAConst1Const2, which does not get significant in this model. LSAConst1Const2 was rather found in interaction with Const1ResidFamSize.

Interestingly, Const1ResidFamSize and LSAConst1Const2 are only significant if they occur in an interaction with each other, which reduces AIC by 80 units. If LSACompConst2 is not included in the model, the interaction does not become significant ($p \approx 0.07$). In Figure 5.13 we see that their interaction functions similar to an AND gate - only if both have high values the fixation is more at the beginning of the word. Additionally, a random forest was fitted to the first fixation position and it showed LSAConst1Const2 as the fifth most important of all 16 predictors.

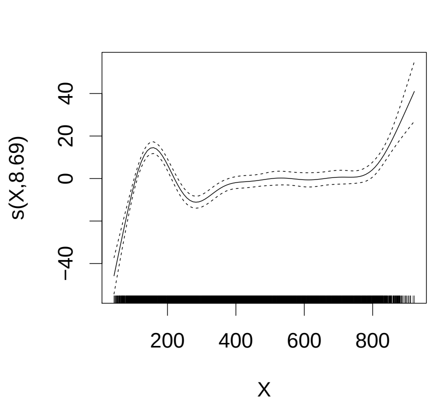

Figure 5.10: Smooth partial effect of X on first fixation position.

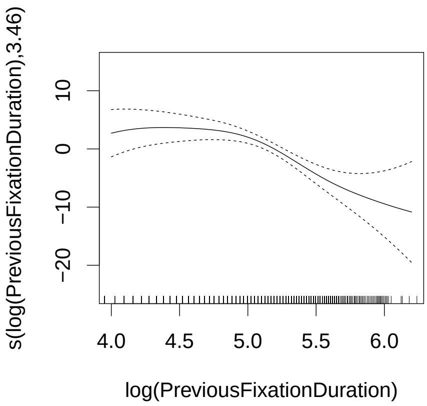

Figure 5.11: Smooth partial effect of PreviousFixationDuration on first fixation position.

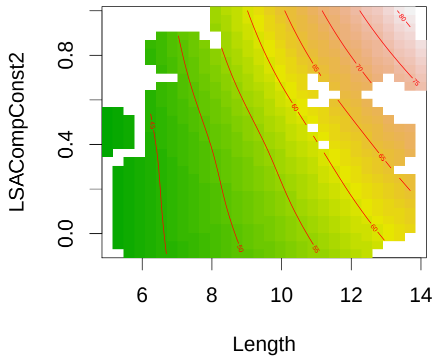

Figure 5.12: Tensor product smooths for the interaction of length by whole-right LSA similarity for first fixation positions of non-space separated compounds.

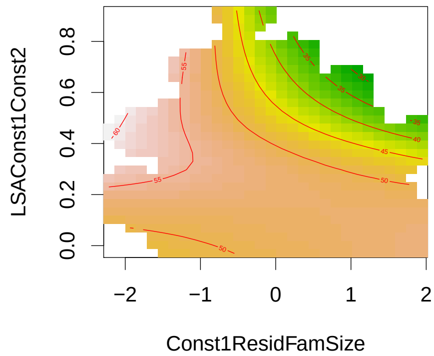

Figure 5.13: Tensor product smooths for the interaction of first constituent family size by left-right LSA similarity for first fixation positions.

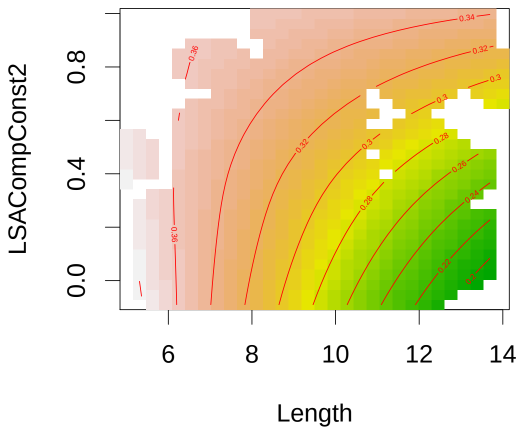

Figure 5.14: Tensor product smooths for the interaction of compound length by whole-right LSA similarity for percentage first fixation positions of non-space separated compounds.

6 General Discussion

In general, reading is strongly governed by individual subject and words differences as indicated by random intercepts which almost always became highly significant. When reading over multiple lines, the position in the line has substantial effects on the duration of the fixation and whether a refixation has to be made. In fact, the horizontal position has come out as the second most important predictor for both first fixation duration and first fixation position in random forests. This is probably because moving from one line to the other and the inherent visual search for a starting point is a dominant factor in the data. It might potentially be harmful to present texts in multiple lines when very subtle effects are to be explored. The most important clue for duration and position is provided by information whether a fixation is single or first-of-many. Single fixations last longer, are on shorter words, more in the center of the word and are influenced differently by frequency measures.

The compound frequency effect was present in first fixations, the first constituent frequency additionally in first-of-many fixations. Thus, despite that we can not safely state that the first constituent occurs in the model of single fixations, we can not falsify hypothesis (1) completely: at least the first constituent and the whole compound affect lexical processing of compound words. We were not able to find an effect of second constituent frequency neither by analyzing single nor last of many duration.

Further, since we observed the time-course of frequency effects and this reflects activation of the mental representations, we can say something about morphological processing. If morphological processing was purely sublexical one would expect first activation of left constituent, then right constituent and then whole compound. On the other hand, if it was purely supralexical we would have to see activation of the compound, then left constituent and then right constituent. Thus, if hypothesis (2) - compound, first constituent frequency and family size precede right constituent frequency and family size - is true, then the strict models are implausible. We are able to confirm this in part, because for first-of-many fixations we witnessed that both frequency of the first constituent and compound frequency show up. Since there is a refixation to be made this reflects early processing and speaks against strict sub- or supralexical processing. Kuperman et al. (2009) reported similar results but their experiment consisted of isolated words instead of words in sentence context, which is important considering possible spillover effects like we saw with previous fixation duration. Therefore, it would be reasonable to follow their argument for a parallel dual route or multiple interactive route model.

A model in which a compounds meaning is both retrieved by full-form access and computation from the constituent receives additional evidence due to the role of transparency in fixation durations. Semantic transparency operationalized by LSA similarity between compound and second constituent has a reasonable effect in single fixations. If the compound frequency is high, transparency plays no role because direct lexical access is the fastest pathway. If compound frequency is low, some computation is necessary and its speed is enhanced by transparency. On the other hand, the compound frequency has almost no effect in highly transparent compounds. In earlier measurements, namely in first-of-many fixations, transparency has a facilitating effect only when the fixation was far into the word. This indicates that properties of the second constituent can be identified early. Additionally, longer compounds are are more likely to require a refixation if they are transparent. These results strengthen hypothesis (3), LSA similarity helps predicting the first fixation duration, but from the duration models it does not look like it plays a major role at early stages.

That LSA similarity plays a role at surprisingly early stages is shown in the model for first fixation position and strengthens hypothesis (4). Even though the compound has not yet been in foveal inspection, semantic processes are in full swing, which is consistent with the results showing that processing of upcoming words happens during reading of the current word (Kliegl et al., 2007) and resembles the findings of Baayen et al. (2013). In contrast to their results, a high right-whole LSA similarity leads to fixations more to the right. One advantage we have over their methodology is that we observed fixations at 944 uniquely embedded compounds, whereas they only presented 111 compounds. Baayen et al. (2013) additionally reported that higher left-right LSA similarities result in fixations more to the right, which is not the case in our results. Since higher left-right LSA similarities indicate greater textual interchangeability, it would be reasonable to choose a conservative fixation more in the beginning of the word. The influence of semantic factors speaks strongly for the form-and-meaning account of word processing and enables us to narrow down the time needed for semantic processing. According to Reichle and Reingold (2013) parafoveal processing takes place during saccade programming, i.e. during $80-120$ msec before the beginning of a saccade. This allows for a maximum of 120 msec for initial semantic work, which is way below the 300 msec lower boundary of semantic processing reported in neuroimaging studies.

The reason why we are able to find early effects of semantics is probably due to the usage of eye tracking instead of the eye-contingent boundary paradigm used in Rayner et al. (1986). Additionally, LSA measurements have been found to resemble more closely human performance then human transparency ratings. Instead of isolated words, reading was performed in sentential context which provides a variety of additional cues. Finally, the usage of a regression design and especially GAMs allow for detection of subtle non-linear effects.

Baayen et al. (2011) and Baayen et al. (2013) propose the naive discriminative reader to account for the wide range of morphological effects such as the form-and-meaning results. It is a model in which orthographic letter representations are directly mapped onto semantic representation (lexemes), without specific representations for morphemes or whole words. Rather, it is a bottom-up approach that uses the Rescorla-Wagner equations to define the association strength of a letter trigram to a semantic outcome. To judge the performance of the model, it was used to compute activation levels of the outcomes given the orthography of a word. Then, the activations of compound, modifier and head were utilized as predictors in the GAM models and provided a much tighter fit than the standard predictors we discussed here. This poses a problem for multi-stage models that assume many theoretical constructs such as the dual route model.

It is presumed that the activation spreads to related lexemes from which additional predictions are generated. If the bottom-up process fails, i.e. a critical orthographic cue is missed, top-down processes motivated by the activation of related lexemes are initiated. As such, form and meaning processing appear hand in hand. The lexeme representations are also triggered by events that happen in the world, so it is assumed that the activation spread to semantically related lexemes is mediated by real world cues. Since there is no data for experiential cues, semantic relatedness can only be measured with LSA, which therefore still appears in the GAM models. For the initial fixation position, if the activation levels were included, LSACompConst2 had the same effect as we saw in our experiments: higher LSA similarity leads to further fixations. In terms of uncertainty reduction this makes sense because it allows input from the right of the word to give contrastive evidence and resolve the confusion of head and compound.

The naive discriminatory reader has the advantage that it is a explicitly articulated computational model, as opposed to the often verbal models. It is a form-and-meaning approach, that explains the observed effects of compound processing with uncertainty reduction. Therefore, even without computing activation levels and testing the performance in GAMs it can be safely stated that it is a promising candidate model for language processing.

7 References

Anderson, S. R. (1992). *A-Morphous Morphology.* Cambridge University Press.

Baayen, R. H. (2010). A real experiment is a factorial experiment? *The Mental Lexicon*, 5(1):149–157.

Baayen, R. H. (2011). *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics".*

Baayen, R. H., Dijkstra, T., and Schreuder, R. (1997). Singulars and Plurals in Dutch: Evidence for a Parallel Dual-Route Model. *Journal of Memory and Language*, 37(1):94–117.

Baayen, R. H., Kuperman, V., Ramscar, M., and Shaoul, C. (2013). Decomposition makes things worse: A discrimination learning approach to the time course of understanding compounds in reading.

Baayen, R. H. and Milin, P. (2010). Analyzing Reaction Times. *Marketing Science*, 12(4):395–414.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, 118(3):438–81.

Baayen, R. H., Piepenbrock, R., and Van Rijn, H. (1993). *The CELEX lexical database (CD-ROM)*, volume 358.

Bertram, R. (2011). Eye movements and morphological processing in reading. *The Mental Lexicon*, 6(1):83–109.

Bertram, R., Baayen, R. H., and Schreuder, R. (2000). Effects of Family Size for Complex Words. *Journal of Memory and Language*, 42(3):390–405.

Booij, G. (2005). *The grammar of words: an introduction to linguistic morphology.*

Brants, T. and Franz, A. (2006). Web 1t 5-gram version 1. Technical report, Linguistic Data Consortium, Philadelphia.

Brysbaert, M. and New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4):977–90.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7(3):249–253.

Feldman, L. B., O'Connor, P. A., and Del Prado Martín, F. M. (2009). Early morphological processing is morphosemantic and not simply morpho-orthographic: a violation of form-then-meaning accounts of word recognition. *Psychonomic bulletin review*, 16:684–691.

Forster, I. and Taft, M. (1976). Lexical Storage and Retrieval and Polysyllabic. *Journal of Verbal Learning and Verbal Behavior*, 15(6):607—-620.

Francis, W. N. and Kucera, H. (1979). Brown corpus manual. *Brown University Department of Linguistics*.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Giraudo, H. and Grainger, J. (2001). Priming complex words: evidence for supralexical representation of morphology. *Psychonomic bulletin & review*, 8(1):127–31.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical science*, pages 297—-310.

Hohenstein, S., Laubrock, J., and Kliegl, R. (2010). Universität Potsdam Semantic preview benefit in eye movements during reading : a parafoveal past-priming study. *Jounal of Experimental Psychology: Learning, Memory, and Cognition*, 36(5):1150–1170.

Hyönä, J., Bertram, R., and Pollatsek, A. (2004). Are long compound words identified serially via their constituents? Evidence from an eye-movement-contingent display change study. *Memory & cognition*, 32(4):523–32.

Kliegl, R., Risse, S., and Laubrock, J. (2007). Preview benefit and parafoveal-on-foveal effects from word n + 2. *Journal of experimental psychology. Human perception and performance*, 33(5):1250–5.

Kuperman, V., Schreuder, R., Bertram, R., and Baayen, R. H. (2009). Reading polymorphemic Dutch compounds: toward a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, 35(3):876–895.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 1(2):211–240.

Lavric, A., Clapp, A., and Rastle, K. (2007). ERP evidence of morphological analysis from orthography: a masked priming study. *Journal of cognitive neuroscience*, 19(5):866–77.

Marelli, M. and Luzzatti, C. (2012). Frequency Effects in the Processing of Italian Nominal Compounds: Modulation of Headedness and Semantic Transparency. *Journal of Memory and Language*, 66:644–664.

Morrison, R. E. (1984). Manipulation of stimulus onset delay in reading: evidence for parallel programming of saccades. *Journal of Experimental Psychology: Human Perception and Performance*, 10:667–682.

R Development Core Team, R. (2011). R: A Language and Environment for Statistical Computing.

Rastle, K. and Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language & Cognitive Processes*, 23:942–971.

Rastle, K., Davis, M. H., and New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, 11(6):1090–1098.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Rayner, K., Balota, D. a., and Pollatsek, a. (1986). Against parafoveal semantic preprocessing during eye fixations in reading. *Canadian journal of psychology*, 40(4):473–83.

Reichle, E. D. and Reingold, E. M. (2013). Neurophysiological Constraints on the Eye-Mind Link. *Frontiers in Human Neuroscience*.

Schreuder, R. and Baayen, R. H. (1995). Modeling morphological processing. In *Morphological aspects of language processing*, pages 131–154.

Seidenberg, M. and Gonnerman, L. (2000). Explaining derivational morphology as the convergence of codes. *Trends in cognitive sciences*, 4(9):353–361.

Seidenberg, M. S. and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523–68.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307.

Taft, M. and Forster, K. (1975). Lexical storage and retrieval of prefixed words. *Journal of verbal learning and verbal behavior.*

von Humboldt, C. W. F. (1836). *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwicklung des Menschengeschlechts.* Dr. d. Kgl. Akad. d. Wiss.

Williams, E. (1981). On the notions" Lexically related" and" Head of a word". *Linguistic inquiry*, 12(2):245—-274.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*, volume 170.

Xie, Y. (2011). knitr : A General-Purpose Tool for Dynamic Report Generation in R. *Design*, 8:1–11.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Bachelorarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Ort, Datum | Unterschrift