Zipf’s Law in Aphasic Speech

An Investigation of
Word Frequency Distributions

Published by
LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6111

email: lot@uu.nl
http://www.lotschool.nl

Cover illustration: created by Marjolein van Egmond

ISBN: 978-94-6093-298-4
NUR 616

Zipf's Law in Aphasic Speech

An Investigation of
Word Frequency Distributions

De wet van Zipf in afatische spraak

Een onderzoek naar
de distributie van woordfrequenties

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 26 oktober 2018 des middags te 12.45 uur

door

Marjolein van Egmond

geboren op 22 december 1986 te Gouda

Promotor: Prof. dr. S. Avrutin

Copromotor: Dr. A. Dimitriadis

“If your life’s work can be accomplished in your lifetime, you’re not thinking big enough.”

Wes Jackson

Contents

Acknowledgments

Here it is: my dissertation. Somehow, I did it. I reached the finish line. Along the way, there were quite a few people who helped me do it, encouraged me or were simply there for me. It is my pleasure to thank you all.

The most important one, all way through, was my promotor and supervisor Sergey Avrutin. Sergey, I don’t think I would have finished if it weren’t for you. You never stopped believing in me, no matter how long it took and no matter how much I doubted myself. You never said ‘this doesn’t make sense’ (not even when it really didn’t make sense), always ‘I don’t quite understand this part’. Not only did you support me on the academic level, you were also genuinely happy for me when I got pregnant… twice – not an obvious thing in academics, I have been told. I feel very fortunate that I had you as my supervisor. Thank you for everything.

The next person to thank is my co-promotor and second supervisor Alexis Dimitriadis. Alexis, thank you for joining us in the last year. You provided the structure that was needed to finish this project. I’m sure I would have needed at least another year without you. Thank you for your critical and sharp look. You helped me to tie everything together. I needed that. Thank you.

Lizet van Ewijk, you were my predecessor, my mentor and my example. Thank you for answering all my questions, for helping me with both theoretical and practical problems, and for being a great example on how to live life in general.

This project would not have existed without the participants in this study, both with and without aphasia. I literally could not have done it without you. Thank you for talking to me and for sharing your personal stories with me. You made it worthwile and made me remember why I started this research project in the first place. A big thank you also to the speech therapists at Via Reva, Sophia Revalidatie and Libra Revalidatie & Audiologie for helping me recruit the aphasic participants.

Christian Bentz, thank you so much for allowing me to use your R code!

Over the years, there were a number of students who took part in my project or worked on their own related projects under my supervision. I really enjoyed the discussions with every one of you. Kyriaki, the research presented in Chapter 5 was carried out together with you. Thank you for your dedication to this project, and for all the valuable discussions we had about Zipf's law and aphasia. I thoroughly enjoyed having you on the team. Juan Miguel Vicente Flores, thank you for your interesting work on Zipf's law in L1 attrition. Your passion for this project (despite the many challenges) helped me to re-find my own. Marten Postma and Robert Grimm, thank you for your technical insights and all our discussions on that impossible question about when Zipf's law does *not* apply. Marten, thank you for your help on previous versions of the growth curves. Nadine Roos, thank you for your help with the transcription of the speech samples.

Loes van Bree and Desiree Capel, thank you for having my back on the day of my defense as my two paranymfs. Loes, thank you for being there, always, no matter how far apart we are. We've shared half our lifes now. You are not just a friend, you are like an extra sister. Desiree, together with Anna Sara Rømoren and Assaf Toledo we had the best office in Trans. Of all people I know, you know best what I was going through while finishing a marathon dissertation. Thank you for your company, your amusing little noises while you worked, for making age differences disappear and for remembering me when I moved away.

I have many good memories of the Trans and all the people who were there during my many years in Utrecht. I felt supported and altoghether in the right place. There are many people to thank for this. Desiree Capel, Anna Sara Romøren and Assaf Toledo, thank you for making my days brighter in 0.45 and beyond. Assaf, thank you for the beautiful pictures, both of Utrecht and – nowadays – of Israel and Nelly. Keep them coming! Anna Sara, thank you for being there, for listening, talking and sharing coffees with me. Together, let us never forget Paula Cox and her beautiful drawings that were all over the place.
Heidi Klockman and Shuangshuang Hu, thank you for being awesome roommates in the attic in my last few months at Trans. Please never grow up!
Anne van Leeuwen, Eva Poortman, Ileana Grama, Maartje Schulpen, Mirjam Hachem, Carolien van den Hazelkamp, Louise Nell and Anne-France Pinget, we started our projects (pretty much) at the same time. Thank you for making this journey together with me. Thanks to you, it was less lonely.
In no particular order: thank you Stavroula Alexandropoulou, Myrthe Bergstra, Dominique Blok, Renske Bouwer, Bert Le Bruyn, Xiaoli Dong, Lotte Dros-Hendriks, Anja Goldschmidt, Alexia Guerra, Mengru Han, Björn 't Hart (thank you for feeding the pregnant lady!), Marko Hladnik, Brigitta Keij, Suzanne Kleijn, Choonkyu Lee, Liquan Liu, Zenghui Liu, Sophia Manika, Sandrien van Ommen, Liv Persson, Jolien Scholten, Marijn Struiksma, Silvia Radulescu (I wish we had had more research-related discussions than we did), Hanna de Vries, Franca Wesseling, and Rob Zwitserlood, for the lunches at the law faculty, the garden at Trans and elsewhere, the Uiltjesdagen, the coffee, the talks and the feeling of

support and being at home. Maaike Schoorlemmer, thank you for your help with all practical matters one could possibly think of and your genuine interest in my wellbeing.
Anne van Leeuwen, Carolien van den Hazelkamp, Marijn Struiksma and Liv Persson, you were the ones who sparked my love of cycling. My first hesitant kilometres on a racing bike were together with you, and they were followed by many, many more. Thank you for that!

In addition to all work-related people who helped me along the way, there were a number of other important people without whom I wouldn’t have been able to persevere. Bettina Zevenbergen and Patrick van der Hofstad, thank you for teaching me how to meditate. It has helped me in every aspect of my life. It kept me sane, and still does. Saskia, thank you for your practical support and also for your presence. You made working from home less lonely. Susanne Spiele, thank you for listening to me and laughing with me. The distance disappears when I have you on the line. Lieke Klarenbeek, thank you for your support in motherhood, it is good to be able to share this adventure with you. Marije Franken, thank you for making me feel more at home in Bergen op Zoom.

Dearest in-laws Cor and José, Jos and Rémon, thank you for allowing me to be part of your family. It feels good to have you near us. José, thank you for taking care of Koen, all those Tuesdays and more. I love seeing the close bond you have together.

A big thank you to my little brother Steven and little sister Lysanne. You are the best brother and sister I could wish for. Can we please re-organize the Netherlands and put Noord-Holland next to West-Brabant? Mandy and Jochem, I’m happy you joined our family. Thank you for bringing the best nieces ever into my life, Rosemarijn, Linde and Yuna.
Mam and Arie, papa and Nelleke, thank you for everything. Thank you for all the moral and practical support, even though you barely knew what I was doing. Thank you for lifting me up *and* for keeping my feet on the ground. Mam, thank you for always giving me a place to come home to, no matter where that place is.

Most important people always come last. Koen and Tom, thank you for bringing such an enormous amount of love and joy into my life. Your arrival put everything in the right perspective. Koen, you were the best reason for delay I could possibly think of. And Tom, you were the best motivation to finally finish it.
Daan, there is no way I could ever write down what you mean to me in just a few lines. You’ve been there for me during almost all of my adult life. Thank you for supporting me in every possible way. Thank you for putting up with me when I wasn’t my best self because of this project. Thank you for being in my life. I love you.

1 General introduction

1.1 A curious pattern

Word frequencies in a text – at least in any written text – follow a curious pattern. A few of them appear extremely frequently, while by far most of them appear only once or twice. In fact, about half of the tokens in a text are comprised of just one or two very frequent types, while at the same time about half of all types occurs only once. This is the case not just in English or Dutch, but in any natural language examined so far. This pattern is considered a law of word frequencies, and better known as *Zipf's law*.

Its universality renders Zipf's law one of the very few true language universals (e.g. Montemurro & Zanette, 2011; Jäger & Van Rooij, 2007). The discovery of Zipf's law is not a new one: its existence has been known for more than a century, at least since 1916 (Estoup, 1916). And yet it is still largely covered in a veil of mystery. It has extensively been studied in large corpora of written texts and in many different languages (e.g. Baayen, 2001; Balasubrahmanyan and Naranan, 2002; Baroni, 2008; Baroni, Bernardini, Comastri, Piccioni, Volpi, Aston & Mazzoleni, 2004; Baroni & Ueyama, 2006; Popescu, Altmann & Köhler, 2010; Zipf, 1949). But that is about it. There is no way to say when it does *not* hold. We hardly know how it behaves in more singular cases such as language output from people with language disorders, or even in the not so special case of spoken language. We know little about its behaviour in small text samples. These cases are usually discarded beforehand as not suitable for the study of Zipf's law. But if Zipf's law is a true language universal that holds for all texts, then it should in theory also hold for small texts and spoken texts (albeit perhaps with different parameter values).

In the case of language disorders, it is less clear what to expect. Most likely, it depends on the nature and severity of the disorder under consideration whether Zipf's law holds and if so, with which parameter values. But Zipf's law is not (only) an object of study on its own, it can also fruitfully be applied to study other language phenomena (e.g. Yang, 2010, for child language). Such an approach might aid in discovering more about the disorder and the workings of the human language faculty. For this application, too, it is necessary to have insight into its exact workings.

The overarching goal of the current study is to learn more about the behaviour of Zipf's law in these kinds of understudied situations. Zipf's law in the margins, so to speak. The margins under consideration are spoken language, small samples and speech from people with the acquired language disorder aphasia.

1.2 Structure of this dissertation

Before diving into the analysis of different kinds of texts, some theoretical foundations need to be laid. This will be done in Chapter 2. The first is obviously an introduction to Zipf's law. What is it exactly, and why is it there? How did others explain its existence? The explanations for Zipf's law fall into a number of categories: The Principle of Least Effort, intermittent silence, preferential attachment, new optimization models (present day versions of the Principle of Least Effort) and a variety of semantic explanations. I will discuss the reasoning of each approach, and the plausibility of it.

The other necessary foundation is an introduction into the internal process of speech production, in combination with a discussion of the acquired language disorder aphasia. The important concepts of age of acquisition and word frequency will also be discussed.

Zipf's law holds for many different texts from many different sources and languages. When it holds, word frequencies plotted against rank (most frequent word on rank 1, second most frequent word on rank 2, etc) follow a log-linearly declining trajectory. One of the parameters of the formula, parameter α, reflects the slope of this trajectory (see further Section 2.2.1). But the many different values that were found for this slope of the word frequency distribution throughout the literature cannot directly be compared because of methodological differences between the studies. It therefore remains unclear if any systematic differences exist between texts from different sources. These questions will be addressed in the first part of Chapter 3. The main question that is relevant for the current study is whether any systematic differences exist between the frequency distributions of written and spoken language. Therefore, a continuum is constructed from spontaneous, spoken language (free conversation) to heavily edited written language (novels), in 7 steps. Texts in each category are analysed for Zipf's law. Identical methods now allow for a direct comparison of text types.

It is also unclear what a reasonable lower limit would be for the number of words that should be included in any analysis of Zipf's law. Usually, it is claimed that the longer the better, but in the case of spoken language, long samples are not always available. Growth curves for the values of Zipf's law are constructed for the texts in the different categories from spoken to written text, allowing for an answer to this question.

The second part of Chapter 3 takes a very different approach to explore the workings of Zipf’s law. The fact that Zipf’s law is found in every text in every language means that there are no naturally occurring texts for which Zipf’s law does *not* hold. It is therefore an open question if it is at all possible to remove Zipf’s law from a text (while keeping the text readable), and what the result of that would be. How would readers evaluate such a text? These questions are addressed in the second part of Chapter 3.

After the theoretical discussion of Zipf's law in Chapter 2 and the more practical exploration of Zipf's law in Chapter 3, the stage is set for a study of Zipf's law in speech from people with the language disorder aphasia in Chapter 4. Are there any differences between non-fluent aphasic speakers and healthy controls in terms of Zipf's law?

In Chapter 2 it is discussed that some hypotheses for Zipf's law relate its existence to the organization of the human lexicon, which is thought to be a network of related words and concepts. This network is thought to be structured based on the age at which the words in it are acquired when a child learns the language (Age of Acquisition, AoA) (Steyvers & Tenenbaum, 2005).[1] Age of acquisition itself is highly correlated to general frequency in the language under consideration. For that reason, AoA and frequency are here studied in the speech of people with aphasia as well. Any aberrations in the distribution of Zipf's law are expected to be correlated with aberrations in the distributions of AoA and general word frequency. To my knowledge, this is the first time that AoA and general word frequency are studied in samples of spontaneous speech.

The speech from people with aphasia is not only studied at one single point in time, but rather participants were interviewed at three different moments after the onset of their aphasia. All participants were interviewed at 2, 5 and 8 months post onset, the period of time in which patients go from the acute stage of aphasia to the chronic stage. Can any developments in the measures under consideration be seen when patients recover? Do any measures correlate with the severity of the disorder? These are the questions addressed in Chapter 4.

[1] I would like to stress that I do not refer to the age at which these particular participants acquired the words they use, but rather about the age at which the words they use are usually acquired by average speakers of Dutch.

The participants from Chapter 4 all spoke Dutch. In Chapter 5, it is tested to what extent the non-fluent aphasic speech findings can be generalized to other languages, namely English, Greek and Hungarian. English is – just as Dutch – a morphologically simple language, without much inflection. Greek and Hungarian, on the other hand, are considered morphologically complex languages. It is tested whether any differences exist between these different languages. In addition, it is explored how Zipf’s law behaves in fluent aphasic speech in English, Greek and Hungarian. To my knowledge, this is the first time that Zipf's law is studied in fluent aphasic speech.

The speech samples that were available in these additional languages were rather short, much shorter than what is usually considered a large enough sample to studied Zipf's law. In addition, therefore, a method is presented to be able to study Zipf's law in small samples, by comparing the outcomes of the test groups to large baseline corpora of speech samples from healthy speakers. The possibility to use small-sized samples opens up the possibility to study Zipf's law in all sorts of populations for which long speech samples are not available.

Finally, in Chapter 6 a General Discussion and Conclusion is presented in which I reflect on the studies in the previous chapters and discuss to what extent we now know more about the workings of Zipf's law. Did we get any closer to knowing where it originates? In addition, did we gain insight into the language faculty of people with aphasia? It is my wish that this study fills some gaps in the knowledge in both fields.

1.3 Research questions

To summarize, the research questions for this dissertation are as follows:

With respect to Zipf's law in general:

1. What are the current hypotheses for Zipf's law, and which one seems to have most potential?
2. How does Zipf's law behave for very small text sizes, and is there a lower limit in terms of number of tokens below which Zipf's law does not hold?
3. Are there any systematic differences in terms of Zipf's law between spoken and written texts?
4. a. Is it possible to create a text for which Zipf's law does not hold?
   b. How do readers evaluate such a text?

With respect to Zipf's law and aphasia:

5. a. Are there any systematic differences in terms of Zipf's law between people with and without non-fluent aphasia?
   b. Is there a difference in the distributions of
      - AoA and/or
      - frequency in the Dutch language in general

      in the speech of people with and without non-fluent aphasia?
   c. In addition, does any of the measures above correlate with the severity of non-fluent aphasia, and do we see any developments when people recover?
   d. Does Bentz and colleagues' NFD measure (as proposed in Bentz et al., 2017, and discussed in Chapter 4, Section 4.1.2) add anything to the picture as it follows from the questions above?
6. Can any differences found between Dutch healthy and aphasic speakers be generalized to other languages?
7. Are there any systematic differences between fluent and non-fluent aphasic speakers?
8. Are there any possible clinical implications or possibilities for people with aphasia or other impairments that follow from the study of Zipf's law?

The reason for this study is twofold.

My hope is that someday, we will be able to pinpoint exactly where Zipf's law comes from. Is it only a statistical regularity, as some claim, or is it the result of something deeper, such as the organization of the human language faculty? We can only hope to answer that question if we know more about its behaviour in texts that are a little more exotic than large, written corpora. This will be my contribution to the field of Zipf's law, written down in the dissertation that you are reading now.

But in my opinion, Zipf's law is much more interesting if it is not only used as an object of study in itself, but also as a tool to gain more insight into the workings of language in a more general sense. My ultimate hope is that my research can somehow help people with a language disorder. With this in mind, I attempt to extend the use of Zipf's law to the field of aphasiology. Zipf's law might prove to be useful to study this complex disorder from a point of view that is independent from any linguistic theory. I hope that either Zipf's law can be used to classify people with aphasia, or that through Zipf's law, I can help to provide insight into the workings of the lexicon in people with aphasia, insight which then might be used to develop or refine recovery programs.

2 Theoretic framework

For a good understanding of the studies reported on in this dissertation, two foundations have to be laid. One is an introduction to Zipf's law. What is it, and why is it there? The existence of this law has been known for a long time. How did others explain its existence? The explanations for Zipf's law fall into a number of categories: The Principle of Least Effort, intermittent silence, preferential attachment, new optimization models (present day versions of the Principle of Least Effort) and a variety of semantic explanations. I will discuss the reasoning of each approach, and the plausibility of it. The other necessary foundation is an introduction into the process of speech production, in combination with a discussion of the acquired language disorder aphasia. The important concepts of age of acquisition and word frequency will also be discussed.

Some hypotheses for the existence Zipf's law refer to the system of word storage and speech production. For this reason, I will start with this foundation, followed by the explanations for Zipf's law. Together, these foundations will set the stage for the rest of the chapters in this dissertation.

2.1 Speech production

In this dissertation, I am specifically interested in Zipf's law in spoken language. That means that I need a theoretical framework concerning language production. I here choose to use Levelt's model as a basis. Levelt's model is not the only one, there are quite a few more. There are models that use an information-processing framework (the group to which Levelt also belongs), such as Fromkin (1971, Garret (1975, 1976), logogen models, such as Morton (1970) and Morton and Patterson (1980), localist connectionist models such as Dell (1986; Dell & O'Seaghdha, 1992), Harley (1984), Harley & MacAndrew (1992), Stemberger (1985) and Rapp & Goldrick (2000), and distributed connectionist models of word processing, such as Plaut & Shallice (1993a, 1993b) and Plaut (1996) (for a discussion of these different models, see Laine and Martin, 2006, p. 15-35). They often share many of the steps in the process, but few are developed into such detail as Levelt's model is. This model, therefore, promises to be most helpful for my research.

2.1.1 Levelt's model

Levelt's model consists of a number of stages, each producing its own characteristic output representation. The first two stages are part of the conceptual/syntactic domain, the other three are part of the phonological/articulatory domain. The following discussion of this model is based on the version as described in Levelt, Roelofs and Meyer (1999), depicted in Figure 2.1.

Word production starts with the stage of *conceptual preparation*. The output of this phase consists of lexical concepts. There is no simple one-to-one mapping of notions-to-be-expressed onto messages. Instead, the speaker can use a specific term (e.g. 'foal'), a more generic term (e.g. 'animal') or a phrase (e.g. 'baby horse'), dependent on the specific circumstances. This is called *perspective taking*. Lexical concepts are thought to form a network of connected concepts, in which spreading of activation takes place from one activated concept to a connected next one.
The lexical concepts feed into the phase of *lexical selection*. Lemmas are selected that are needed to express these lexical concepts. This is done based on spreading activation from the lexical concepts to the connected lemmas, in a statistical kind of way: the lemma with the highest level of activation is retrieved. Upon selection of a lemma, its syntax becomes available for further grammatical encoding, allowing for example for number- and gender-marking.

The next step in the process is *morphological encoding and syllabification*. We now move from the conceptual/syntactic domain into the phonological/articulatory domain. The first step in this is to take the selected lemmas and retrieve their phonological shape from the mental lexicon. This shape consists of three kinds of information: its morphological makeup (the individual morphemes that together form the word), its metrical shape (stress placement) and its segmental makeup (the individual phonemes of the word). The result of this step consists of the phonological word. This phonological word is input to the stage of *phonetic encoding*, which will result in the words phonetic gestural score, likely through the activation of complete syllables in a syllabary, a repository of gestural scores for the frequently used syllables of a language. The gestures to be performed are coded quite abstractly (e.g. 'close the lips' to produce a word as 'apple'). The final step in the process is the actual *articulation*, when the word's gestural score is executed by the articulatory system.

These five steps in the process of word production are thought to function through spreading activation in three different strata of nodes. The first stratum is the conceptual one, containing concept nodes and labelled conceptual links. A subset of

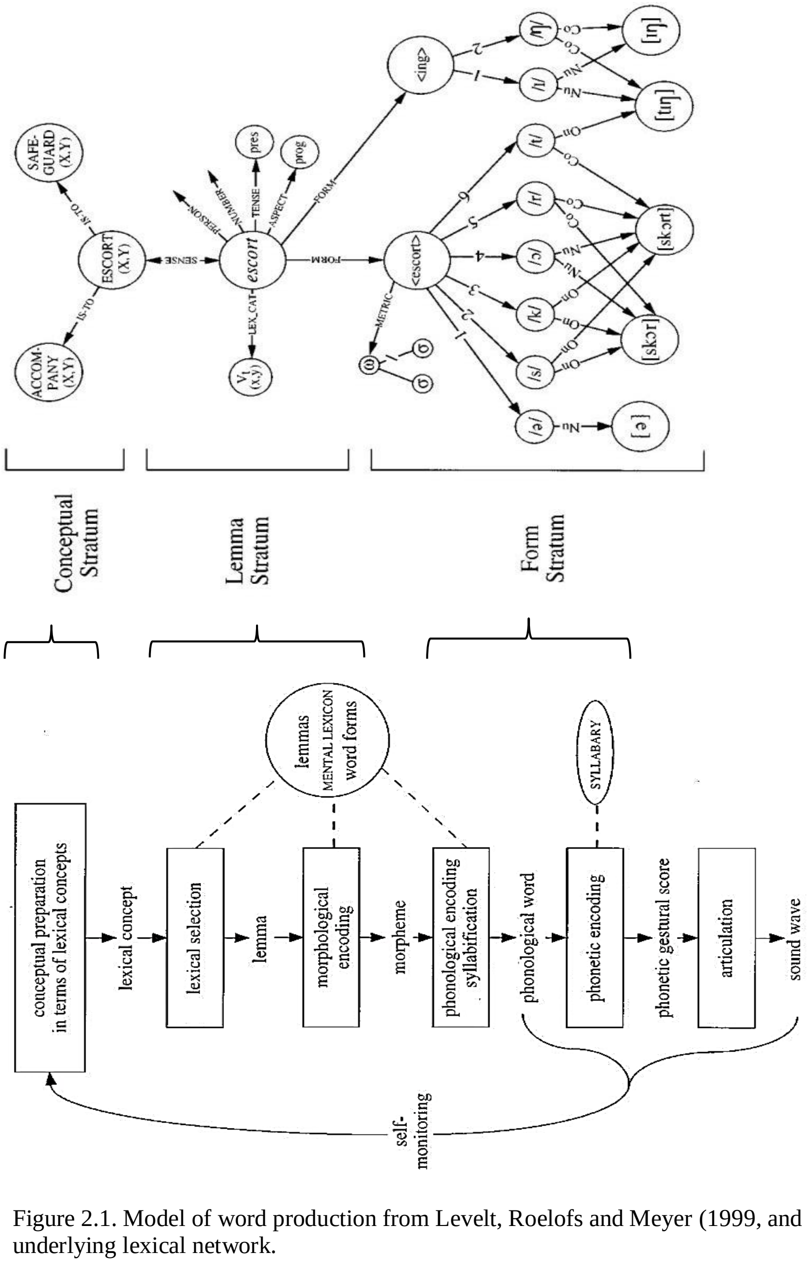


Figure 2.1. Model of word production from Levelt, Roelofs and Meyer (1999, and the underlying lexical network.

these nodes are lexical concepts, which have direct links to lemma nodes in the next stratum. A word's meaning is represented by the total of the lexical concept's labelled links to other concept nodes. The second stratum consists of the lemma nodes, containing the syntactic properties (e.g. $V_t(x,y)$ for a transitive verb with agents x and y) and labelled links between them. Each word in the mental lexicon is represented by a lemma node, it's syntax is represented by the labelled links of its lemma to the syntax nodes.
Once the lemma is selected, activation spreads to the third stratum which is the word-form stratum. This stratum contains morpheme nodes and segment nodes.
A *lexical entry* in this structure is formed by an item in the mental lexicon, consisting of a lemma, its lexical concept (if any, in case of function words) and its morpheme(s) with their segmental and metrical properties.

Sentences, and as a result, complete texts, are thought to be formed by successive runs through the word production system. The individual stages of word production are being taken at an incredible pace: we usually produce about two to three words per second. Nevertheless, the processing stages are considered to be strictly serial: there is neither parallel processing nor feedback between lexical selection and form encoding (except for the specific case of self-monitoring), although it is possible to use information from other words that came available at intermediate steps such as lemma information from nouns when the correct form of an adjective has to be chosen. There is also no free cascading of activation trough the network, nor do inhibitory links exist anywhere in the system. There is competition, though, and the level of activation of non-target nodes influences the latency of the selection of target nodes. Generally speaking, for any smallest time interval, given that the selection conditions are satisfied, the selection probability of a lemma node equals the ratio of its activation to that of all the other lemma nodes (for the mathematical characteristics, see Roelofs 1992; 1993; 1996; 1997; or Levelt, Roelofs & Meyer, 1999). In other words, it is the level of activation relative to that of other nodes that determines a lemmas availability. This level of activation is determined by a number of factors, such as the amount of activation spreading from the conceptual level (where the relevant perspective was chosen) and recent usage (priming).

The plausibility of the model from Levelt and colleagues for language production has been confirmed by a large number of studies, covering error rates, reaction times and modelling (specifically, the WEAVER++ model) (see Levelt et al., 1999, for an extensive review).

2.1.2 Word frequency effect and AoA effect

In many studies, it was found that words that occur more frequently in language are processed faster, more easily or more accurately than words that occur rarer (see e.g. Bybee & Hopper, 2001 for a thorough introduction). This so-called frequency effect is correlated or related to the Age of Acquisition effect, the finding that words that are (estimated to be) learned early in life are processed faster, more easily or more accurately. Especially the concept of age of acquisition will turn out to be of relevance for the discussion of Zipf's law. It is therefore worthwhile to look into these factors in some more detail. Where in Levelt's model could these well-known effects originate, and how can they be explained?

Word frequency effect

I start with the frequency effect. The conclusions from different studies attempting to locate the frequency effect in the steps of Levelt's model are diverse. One highly influential study found evidence for a frequency effect at the phonological level only. Jescheniak and Levelt (1994) had subjects translate various high and low frequency words, including homophones such as *more* and *moor*. The critical question was whether low-frequency *moor* would behave like other, nonhomophonous, low-frequency words (e.g. *marsh*), or rather like other, nonhomophonous high-frequency words (e.g. *much*, matched in frequency to *more*). What they found was that the low-frequency homophones (such as *moor*) were processed statistically as fast as the high-frequency controls (such as *much*). These words thus inherit the fast access speed of their high-frequency partners. This suggests that the frequency effect must originate in accessing the word form rather than the lemma. This finding is in line with Nickels' (1995) data concerning errors from people with aphasia, who found an effect of frequency for phonological errors but not for semantic errors.

Other studies, however, had contradicting results. Caramazza, Costa, Miozzo and Bi (2001) found no evidence that pictures of low-frequency homophones were named faster than those of other low-frequency words. And Jescheniak, Meyer and Levelt (2003) found that low frequency homophones were translated faster than other low-frequency words, but not as fast as high-frequency homophones. In addition, there were studies finding frequency effects on semantic errors. Feyereisen, Van der Borght and Seron (1988), for instance, found that people with aphasia produced fewer semantic paraphasias in the production of more frequent nouns. Nickels and Howard (1994) had mixed results. They studied semantic errors in 15 people with aphasia. The majority of them fail to show a frequency effect in the production of semantic errors. Two of them, however, display an interaction effect between frequency and imageability: they produce fewer semantic errors for words with high as opposed to low frequency, but only for low imageability items.

One more interesting study in this respect is the large-scale regression analysis of aphasic picture-naming errors performed by Kittredge, Dell, Verkuilen and Schwartz (2008). They included error data from a diverse group of 50 people with aphasia and studied multiple variables that might influence word production, amongst which word frequency. The word production model they applied was an interactive (allowing feedback), simplified version of Levelt's model and consisted of two parts, one for lexical retrieval and one for phonological retrieval. Conceptual planning was left out of the model. They found that frequency affected both phonological retrieval and lexical retrieval. According to Kittredge and colleagues, this is compatible with theories placing the frequency effect at all levels of word production.

It might be the case that an item gets easier not when it is retrieved more often (which is what a simple frequency effect is), but rather that it matters how often on average different grammatical operations are performed with that specific word. This number is reflected in the information load of the word, an information theoretic measure of word form complexity (e.g. Kostiç, Markovic & Baucal, 2003; Van Ewijk, 2013). Van Ewijk finds that measures of information load outperform frequency measures in predicting response latencies in an auditory lexical decision experiment. This shows that it might be the combination between frequency and linguistic characteristics that affects lexical processing.

Frequency, possibly in combination with the specific linguistic characteristics of words, thus seems to play a role at multiple levels of speech production.

Age of Acquisition effect

The locus of AoA effects is even more disputed. An extensive review on the AoA-effect was provided by Juhasz (2005). Considering the evidence supporting and contradicting the different hypotheses that have been put forward for this effect, Juhasz concludes that the most likely explanations for the AoA-effect are the Semantic Locus Hypothesis, the Network Plasticity Hypothesis and the Lexical-Semantic Competition Hypothesis. I will not repeat Juhasz' review, but only discuss those three hypotheses.

The *Semantic Locus Hypothesis* states that it is the order of acquisition that is most important, because later learned concepts are built upon earlier ones. The AoA-effect would then arise at the step of conceptual planning. This means that this effect should surface in tasks requiring semantic operations. AoA-effects have in fact been found in various semantic tasks (e.g. Brysbaert, Van Wijnendaele & De Deyne, 2000; Ghyselinck, Custers & Brysbaert, 2004, Juhasz & Rayner, 2003). This hypothesis also predicts larger AoA-effects in tasks that involve semantic representations to a greater degree. Juhasz (2005)' review shows that this is indeed the case: the largest AoA-effects are found in picture naming, followed by lexical decision and word naming.

Other support comes from patient studies. Gerhand & Barry (2000) found significantly more semantic errors for late acquired words in a patient with deep dyslexia. The erroneously produced words were acquired later than the target words. And Nickels and Howard (1995) found AoA to be a predictor of semantic but not of phonological errors in a group of people with aphasia.

The *Network Plasticity Hypothesis* states that early acquired words determine the structure of the network itself because they are learned when network plasticity is at its largest. The network loses some of this plasticity when more words are added and structures become more determined. According to this theory, AoA could affect those stages of word production that require access to networks, which are the module of conceptual preparation and all steps that require access to the lexicon. Evidence in favor of this hypothesis stems from a study by Nazir, Decoppet and Aghababian (2003), in which children in Grades 1 to 5 had to perform lexical decision tasks with words they learned either in previous grades or in their present grade. They found that errors to newly learned words increased with grade, while performance on previously learned words did not significantly improve in subsequent grades. This suggests that words that were learned later were never learned as well as early learned words.

The Network Plasticity Hypothesis and the Semantic Locus Hypothesis do not rule each other out. The AoA-effect could be due to the loss of neural plasticity in the conceptual network. Exactly this is what was modelled by Steyvers and Tenenbaum (2005) in their model of network growth, which will be discussed extensively below as one of the preferential attachment hypotheses for Zipf's law. In short, they modelled semantic growth, and found that early acquired words are most influential to and most deeply embedded in the eventual structure of the network.

Other studies simulating AoA effects in networks are for instance McClelland, McNaughton and O'Reilly (1995), Ellis and Lambon Ralph (2000), Smith, Cottrell and Anderson (2001), Monaghan and Ellis (2002) and Zevin and Seidenberg (2002) (see Johnson and Barry, 2006, for a review). They all find that the early learned words configure the network when its plasticity is largest. Later learned words are then added to the existing structure, when the network becomes increasingly rigid.

Finally, according to the *Lexical-Semantic Competition Hypothesis*, both a frequency related and a frequency unrelated AoA-effect exist. The related effect stems from the same learning process, while the unrelated effect stems from competition when the correct lemma for a certain concept must be selected, thus localizing the effect at the module of lexical selection (Brysbaert and Ghyselinck, 2006; Belke et al., 2005). The reason to assume two loci for the AoA effect instead of one stems from the different AoA-effect sizes in different tasks. In word naming, there are small frequency and AoA-effects, in lexical decision both are somewhat larger, whereas in picture naming AoA-effects are much larger than frequency effects. Brysbaert and Ghyselinck thus suggest that there is something different about the AoA-effect in these different tasks.[2]
Brysbaert and Ghyselinck argue that the frequency-related AoA effect is observed in tasks such as word naming and lexical decision, while the frequency independent AoA-effect is observed in picture naming, word associate generation and category-instance generation. Belke and colleagues (2005) therefore propose the lexical-semantic competition hypothesis, suggesting that competition arises when a lemma is selected for a specific concept. Lemmas for early acquired words are stronger competitors for any given concept.

AoA was also one of the predictors in the above mentioned large-scale regression analysis of aphasic picture-naming errors by Kittredge, Dell, Verkuilen and Schwartz (2008). They found that AoA influenced phonological retrieval but not lexical retrieval. They argue that these findings are in line with theories that place the AoA-effect at the level of word form retrieval and/or the conceptual system (which was no part of their model).

To summarize the results on AoA, it seems to be likely that the AoA effect originates in the conceptual module of the language production system or at the interface between conceptual preparation and lexical selection. This could (at least partly) be due to early words shaping the network structure when neural plasticity is largest.

[2] But note that the different kinds of tasks that are discussed consist of both comprehension and production tasks. The same factor can have very different results in comprehension or in production. Tabak et al. (2006; 2010), for example, find a facilitatory effect for inflectional entropy (a complexity measure, based on freqency) in a visual lexical decision experiment, but an inhibitory effect in a naming task.

2.1.3 Aphasia

Brain damage such as a stroke, trauma or tumor can result in aphasia, an acquired language disorder. Aphasia is a highly complex condition. The set of symptoms is diverse and slightly different for each patient, depending on the exact part of the brain and the language faculty that is affected.

Patients are often classified according to the classic, syndrome-based taxonomy, based on the work of Pierre Paul Broca (1824-1880), Carl Wernicke (1848-1905) and Ludwich Lichtheim (1845-1928). Broca's aphasia (characterized by effortful, telegraphic style speech) and Wernicke's aphasia (characterized by grammatically correct, fluent but incoherent speech, often containing jargon) are the most well-known types of aphasia according to this taxonomy (Goodglass, 1993, Laine & Martin, 2006). Recently, however, this taxonomy is more and more being questioned. Heterogeneity within subgroups is large whilst there are many similarities between patients in different subgroups (e.g. Cuetos Vega et al., 2010). Including only patients that fall within these traditional categories would exclude the large percentage of speakers with mixed types of aphasia and render any conclusions ungeneralizable to the larger group of people with aphasia. The question is, thus, whether it makes sense at all to use this taxonomy for the classification of patients (Van Ewijk, 2013, p. 17-18).

But no matter how patients are classified, there is one problem that is encountered by almost every one of them. This is the problem of word finding difficulties or *anomia*: problems in accessing and retrieving words from the mental lexicon (Goodglass & Wingfield, 1997). This can result in a blockage when no word is retrieved at all, or retrieval of the wrong (often semantically and/or phonologically related) word. Patients can also use other tactics, like description or settling for a word they acknowledge is wrong but semantically close (such as 'seat' instead of 'stool' in the example from patient E.S.T. in Marshall, 1977, see also Ellis, 2006). Other frequent errors from aphasic speakers are the production of neologisms, and several comprehension problems (but comprehension is beyond the scope of this study).

Laine and Martin (2006) distinguish three main kinds of anomia (but one patient can suffer from more than one kind): semantic anomia, word form anomia and disordered phoneme assembly. In terms of Levelt's model, *semantic anomia* is caused by conceptual- and lemma-level deficits. It is usually paired with comprehension difficulties and relatively preserved phonological abilities. The exact nature of the impairment differs per person. The categories of animals, plant life, and artefacts can dissociate from each other, meaning that one category can be severely affected while the others are relatively spared. The level of concreteness/abstractness of the target word can influence lexical processing, as can imageability. Word-class-specific impairments are also a well-known phenomenon. Noun/verb dissociations for example can exist both ways (but they could in some patients be associated to living/non-living dissociations, abstract/concrete dissociations or imageability). A function word/content word dissociation is known to exist in many agrammatic aphasics, who struggle more with function words than with content words. Some anomic patients, however, show the opposite pattern and perform better at function words than at content words. It can even be the case that the deficit is limited to either spoken or written naming.

In all cases, impairments can be due to either permanent loss of information, or due to impaired access to or retrieval from otherwise unchanged representations. The consensus seems to be that both types can occur. Progressive forms of aphasia (such as semantic dementia) appear to involve knowledge loss, while aphasias associated with stroke or other non-progressive impairments (as studied here) are more often thought to involve impaired access to otherwise intact representations.

Laine and Martin (2006)'s category of *word form anomia*, in terms of Levelt's model, is caused by difficulties in accessing the lexeme that specifies the morphological and phonological information of the target word. This kind of impairment is more well-known under the label of *agrammatism*. Inflection might be dissociated from derivation, regulars might be dissociated from irregulars. Frequency of occurrence might obscure any clear-cut boundary between regularly and irregularly inflected forms, because high-frequency regulars are likely to be stored full-form similar to irregulars. It is a kind of impairment that has been studied extensively. There seems to be nothing wrong with the module of morphological encoding per se, but rather reduced resources due to the acquired brain damage result in morphological encoding (more broadly, syntax) no longer being the most economical way to construct a message (Avrutin, 2006). The effect of this is that people with aphasia rely more heavily on discourse to encode part of the message. However, they only do this when this information can be retrieved from the context. Crucially, these constructions are not wrong or ungrammatical and available to healthy speakers too, in the very specific contexts of special registers (e.g. "Maria vertelde Peter een mop. En hij *lachen*!" "Mary told Peter a joke. And he laugh-INF!"). People with aphasia struggle more with tense than with agreement. Agreement is a purely syntactic operation, that can never be replaced by discourse conditions. Tense, however, *can* be provided by the discourse. People with aphasia rely on this circumvention more heavily than people without aphasia, thus frequently omitting tense (but not agreement) from their utterances (see Avrutin, 2006, for more explanations of aphasic behavior in line with this theory).

The module of morphological encoding is thus often (but not always) skipped when possible. Crucially, the step of morphological encoding *can* be taken when no alternative is available, or whenever enough resources were available. The impairment is thus not due to any knowledge being lost.

The third kind of anomia Laine and Martin (2006) distinguish is that of *disordered phoneme assembly*, caused by problems at the module of syllabification and phonological encoding. These patients have no problems with lexical retrieval, but rather struggle with phonologically distorted language output due to problems with the on-line assembly of the phonological output. Longer words are often more difficult.

Levelt's model is capable of accounting for the specific errors made by aphasic speakers through the WEAVER++ implementation (Roelofs 1997; Roelofs & Meyer 1998; Levelt et al., 1999). Substitution errors, for example producing "calf" instead of "cat", can be made because the wrong lemma is selected. The self-monitoring system is able to filter out many errors but is more likely to miss errors that result in real words, especially if the erroneously produced word is semantically close to the target word. When this happens, the lexical concept that is associated to the wrongly produced word is not activated. This means that the speaker who produces the error (or at least the modelled speaker) does not actually *think* of a calf when it produces this word instead of 'cat'. The self-monitoring system cannot attend to all aspects of speech simultaneously: paying more attention to internal speech and speech planning means having less resources available for monitoring of external speech. This explains why people with aphasia do not always hear the mistakes they are making.

To test if Levelt's model can not only explain aphasic word finding difficulties theoretically but also computationally, Laine, Tikkala and Juhola (1998) built a computational model of Levelt's model. They then tested what happens if you change its parameters. Their SLIPNET model consists of two parts: a lexical-semantic network including 27 lemma nodes (10 targets from a picture naming task they performed earlier and their 17 prominent semantic associates), and a phoneme network, including the core set of Finnish phonemes (8 vowels and 15 consonants). The simulation starts with an initial boost of activation given to one of the lemma nodes (the target node), followed by spreading of activation through the lexical-semantic network. The following factors play a role here: the strength of the activation boost (stronger means a higher advantage for the target node), the number of arbitrary time steps allowed for activation spreading (affecting the range of the spreading of activation, set to a fixed number of 2), the decay rate of activation over time, connection strengths between target and related nodes (four levels were used) and noise (to allow for any errors at all). Within this model, they varied two parameters: noise in the lexical-semantic network and/or in the phoneme network, and selection threshold at the lexical semantic level. All other features were kept constant.

Laine et al. show how their simulations are able to account for the four most frequently encountered error types in 10 people with various kinds of aphasia (classified as Wernicke's, Broca's, conduction and anomic aphasics), and the frequency with which they occur, by varying noise and threshold only:

1. *No responses.* Obtained by increasing the threshold values.
2. *Semantic errors.* Obtained by increasing noise at the lexical-semantic level, thus causing the wrong lemma node to be forwarded to the phoneme level.
3. *Neologisms.* Obtained through a combination of increased threshold values and increased noise at the phoneme level.
4. *Other.* This category includes a variety of occasional responses not classifiable into the three other error categories. This included for instance formal paraphasias (phoneme errors that by chance created real words instead of neologisms), or semantic-then-phonological errors, obtained through added noise at both the lexical-semantic and phoneme level.

Usually in studies concerning error-data from healthy speakers, it is found that an above-chance number of mixed errors is found. Mixed errors are formed by a combination of a lexical and a phonological error in the same word. Laine et al. note that their aphasic data does not conform to this finding. Neither is it something their computational model would predict. They explain this as follows. Levelt's model is very strict in terms of direction of flow of activation: it only flows forward, never backward (in other words, it flows from conceptual/syntactic domain to the phonological/articulatory domain, but never the other way around). This assumption might be wrong: it might be the case that there is some feedback happening. The flow of backward activation, however, is weaker to start with. If all activation is weakened, as is likely the case in aphasia, it might be that feedback activation becomes insignificant. This might explain the difference in error patterns between healthy and aphasic speakers. Alternatively, the mixed error effect might be due to so-called environmental intrusions. This could happen for instance because of intrusion of concepts or words which happen to be in the speaker's mind. A third explanation was already given above, namely that the self-monitoring system is more likely to miss errors that result in real words. A phonological error that happens to result in a real word (thus seemingly also a lexical error) is then missed, especially if it is semantically close to the target word.

SLIPNET was less flexible than Levelt's model due to the reduced number of steps. Nevertheless, Laine and colleagues were able to obtain the same types and error rates at a naming task simulation as their (highly diverse) group of aphasic speakers, by twisting only threshold and noise levels. Clearly, they modelled the most relevant part of the network for the kind of data they were working with (which was out-of-context, single word naming data). This finding suggests that the parts that were modelled are indeed the parts where people with aphasia struggle when retrieving individual words. This is in accessing the mental lexicon, and in selecting the right phonemes.

2.1.4 Summary

According to Levelt's model of speech production, speech is produced in seven consecutive, non-overlapping modules. These are conceptual preparation, lexical selection, morphological encoding and syllabification, phonetic encoding and articulation. The first three steps together form the stage of lexical selection, the other two steps form the stage of phonological retrieval. The exact locus of the word frequency effect, the finding that more frequent words are processed faster and/or more accurately, is unclear, but it seems to be the case that frequency plays a role at multiple levels, possibly more pronounced at the level of phonological retrieval. The locus of the AoA effect is even more disputed. The three most important hypotheses are the Semantic Locus Hypothesis that places the effect at the step of conceptual planning, the Network Plasticity Hypothesis that places the effect at all steps that require access to networks (more specifically the lexicon) and the Lexical-Semantic Competition Hypothesis that localizes the effect at the module of lexical selection. These hypotheses are not necessarily mutually exclusive. The Network Plasticity Hypothesis is most defined and has the most psychological and neurological reality to it.

Speech production can be hindered if brain damage results in aphasia. Aphasia is a highly complex condition: the set of symptoms is diverse and slightly different for each patient, dependent on the exact part of the brain and the language faculty that are affected. One frequently encountered problem is that of word finding difficulties or anomia. People with aphasia can suffer from semantic anomia, caused by conceptual- and lemma-level deficits, word form anomia, caused by difficulties in accessing the lexeme, and/or disordered phoneme assembly, caused by problems at the module of syllabification and phonological encoding. The explanatory power of Levelt's model regarding word finding difficulties was confirmed in computational models.

2.2 Zipf’s law

An understanding of the system of speech production can be helpful in interpreting the different hypotheses that have been put forward to explain Zipf’s law. Some hypotheses directly relate to the storage of words. Others explicitly ignore word storage and claim that it is irrelevant.
However, a discussion of the hypotheses for Zipf’s law can only start after a discussion of the formulas of Zipf’s law.

2.2.1 Zipf’s law: The formulas

Word frequencies in natural language texts typically display a characteristic distribution called Zipf’s law. This distribution is found as follows. Count for each word (token) in a text how often it occurs. Sort the words from most to least frequent and assign ranks accordingly: the most frequent word receives rank 1, the second most frequent word receives rank 2, and so on until all words have their own individual ranks. The ordering of words with equal frequencies is irrelevant. Now, plot them with rank on the x-axis and frequency on the y-axis, using logarithmic scales. The result is a straight line with a slope of approximately -1. Fascinatingly, this is the case irrespective of the text and language under consideration. An example of this for the Dutch novel *Karakter* by Bordewijk (1938) is given in Figure 2.2A.

This statistical law of language was probably first noted by the French stenographer J.B. Estoup (Estoup, 1916), who analysed 75 texts with a total of 30.000 words for their word frequencies (without computers!). It was also noticed by Edward Uhler Condon (1928; see also Levelt, 2013, p. 450-451), an assistant professor of physics at Princeton University, who published a note in *Science* on the relation between the frequencies of words and the ranks of these frequencies. But it was not until the extensive studies by G.K. Zipf (Zipf, 1949) that the law became famous, and subsequently got its name.

Zipf’s law is traditionally formulated as the rank frequency distribution

1) $f(w) \approx \frac{C}{r(w)^{\alpha}}$

where $f(w)$ is the frequency of the word with rank $r(w)$ if words are ordered by decreasing frequency and C is a constant that is mainly dependent on text size.

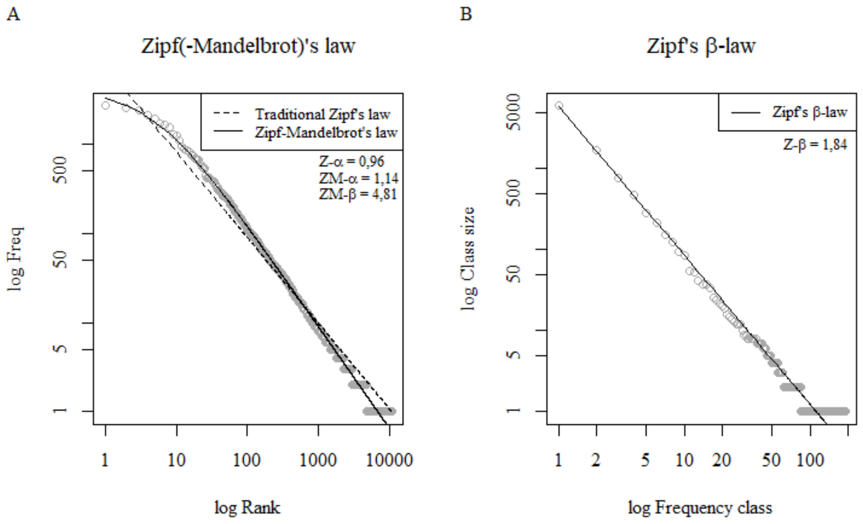


Figure 2.2. (A) The traditional Zipf's law and Zipf-Mandelbrot's law and (B) Zipf's β-law in the novel Karakter by Bordewijk (1983) on logarithmic scales.

The logarithmic version of this law is

2) $\log f(w) \approx \log C - \alpha \log \mathrm{r(w)}$

which directly shows the linear dependency of the variables. The parameter -*α* is now the slope of the distribution.

In the original formulation, it was assumed that $\alpha \approx 1$ (resulting in a negative slope of -1), but nowadays the exponent is assumed to be a parameter and fitted to empirical data, thus allowing it to take on values different from unity (e.g. Zannette & Montemurro, 2005). In fact, substantial deviations from this typical value have been found for different types of texts and for different types of speakers (for a detailed discussion of parameter values of Zipf's law, see Chapter 3).

It has long been known that word frequencies do not follow the straight line of a power law exactly. The most notable difference is that the frequency of high frequency items is overestimated by the power law, resulting in a typical downward curvature for the first few ranks in the plot. To solve this, Mandelbrot (adapted from Mandelbrot, 1954) proposed a modified version of the law:

3) $f(w) \approx \frac{C}{(r(w)+\beta)^{\alpha}}$

This version of Zipf's law is the one used in the current study. The difference in fit for the Dutch novel *Karakter* is displayed in Figure 2.2A. The traditional Zipf's law (Formula 1) is derived when α = 1 and β = 0. Usually, however, β takes on some small value when α is close to 1. It seldom exceeds the value of 10. For small ranks, β is of large influence and decreases the value of $f(w)$. For larger ranks, the influence of β quickly becomes negligible. According to Mandelbrot, β represents the richness of the coding system (in other words, the number of letters in the alphabet) and α depends on the average informativity of a word (measured in entropy, a measure of uncertainty in predicting the word) (see also Section 2.2.3). Others claim that the reason for this typical curvature lies in the different parts of speech that are present in natural language, namely lexical and grammatical words or possibly open and closed class words. Grammatical words are more strongly represented in the higher ranks of Zipf's law, while the lower ranks are mostly filled with lexical words (e.g. Ferrer i Cancho & Solé, 2000; Popescu, Altmann & Köhler, 2010). Both classes of words individually have a different slope. The mixture of the two results in the typical curved shape of Zipf's law.

To achieve better fits to natural language word frequency data, various other variations on Zipf's law and Zipf-Mandelbrots law have been proposed by introducing extra parameters to the function (see Chitashvili & Baayen, 1993, for an overview). However, as Li (2002) puts it: 'It should be pointed out that it is not enough to reject the Zipf's law only because another function fits the data better. The alternative function should not have too many extra parameters in achieving the better fit.' Without an explanation, extra parameters only result in the rather uninteresting exercise of simple line fitting (the same argument is made by Li, Miramontes & Cocho, 2010, p. 1756).

Another version of Zipf's law exists (Zipf, 1949), without the need to rank data. This version is here referred to as Zipf's β-law. The formulation of this version of Zipf's law is as follows:

4) $n_f \approx C \cdot f^{-\beta}$

where the number of words *n* with frequency *f*, in other words, the size of the frequency class, is determined by parameter β (not to be confused with β in formula 3) and a text size dependent constant *C*. This version of Zipf's law has the advantage that it does not depend on the artificial rank variable. This renders it more sensitive to disruptions of the Zipfian distribution: the distribution of Zipf-Mandelbrot's law and the traditional Zipf's law is necessarily decreasing, which is not the case for the distribution of Zipf's β-law.

For texts of sufficient length values on Zipf’s α can be calculated from β and vice versa by using the following formulas (Ferrer i Cancho & Solé, 2001):

5) a. $\alpha = \frac{1}{\beta - 1}$

b. $\beta = \frac{1}{\alpha} + 1$

The traditional version of Zipf’s law (Formula 1), Zipf-Mandelbrot’s law (Formula 3) and Zipf’s β-law (Formula 4) are all formulations of the same phenomenon, and the unspecific term ‘Zipf's law’ has been used throughout the literature to describe them all. In this dissertation, Zipf-Mandelbrot’s law and Zipf’s β-law are the versions that are used for the calculations. To avoid confusion, when either of the two laws is discussed specifically it will be called by its full name. The term ‘Zipf’s law’ is used as an overarching term for all versions of the law.

Traditionally, Zipf's law is the specific instance in which $\alpha \approx 1$ (for the traditional version of Zipf's law and for Zipf-Mandelbrots law) or $\beta \approx 2$ (for Zipf’s β-law). But the name Zipf's law has also been used as the name of the power law that is found in word frequencies, also when that slope is not exactly one and parameters are treated as parameters that can be estimated from the data. This is the use of Zipf's law in the current work too. Another potential point of confusion is the parameter β, because both Zipf-Mandelbrot’s law and Zipf’s β-law have one. Therefore, the β from Zipf-Mandelbrot’s law will hereafter be dubbed ZM-β, while the β from Zipf’s β-law will be dubbed Z-β.

A Zipfian sequence is most easily recognized by its power law behaviour: the first most frequent item is approximately twice as frequent as the second most frequent item, the second most frequent item is approximately twice as frequent as the third most frequent item, etcetera. Zipf’s law is in fact part of a larger class of power laws or scaling laws: heavy-tailed distributions of the form $f(x) \propto x^{\alpha}$ that occur across a range of different fields and phenomena (Kello et al., 2010). When scaling laws are present there are often processes or patterns at work that are repeated across scales of analysis. This means that it does not matter whether you look at the system on a macro or a micro-level; the distribution remains the same (sometimes this requirement holds for part of the system only). This is also referred to as scale-freeness. It is often linked to complex systems in a state close to critical points. At or close to this point, microscopic changes to the system can result in macroscopic effects on a much larger scale (this phenomenon will be discussed in-depth below).

Since its initial discovery, Zipf’s law has been tested numerous times in many different texts and in different languages (e.g. Ha, Stewart, Hanna, & Smith, 2006; Hatzigeorgiu, Mikros, & Carayannis, 2001; Popescu, Altmann, & Köhler, 2010;

Tuzzi, Popescu & Altmann, 2009). For every text in every language, it was shown that Zipf's law holds. Spoken language, although little researched, seems to be no exception (Ridley, 1982; Ridley & Gonzales, 1994). The only possible exception found so far is formed by Simplified Chinese characters, as was shown by Dahui, Menghui and Zengru (2005). They find that mid frequencies are overrepresented, while the number of low frequency items is much smaller than that usually found in Zipf's law. In addition, they find that the total number of tokens is much smaller. The result is a curved slope that drops to zero much faster than that of for instance English. Dahui, Menghui and Zengru claim that this is due to the fact that the number of characters in Simplified Chinese is limited and fixed, whereas in other languages new words can easily be added. They hypothesize that rank statistics of Chinese phrases may conform to Zipf's law (but this was not tested). Diversions from Zipf's law are only found when large enough samples are studied, small samples of Simplified Chinese characters do conform to Zipf's law (Huang, 2014).

Despite its ubiquitous presence, a century after its first discovery still no agreement has been reached about why Zipf's law occurs. Many hypotheses for its existence have been put forward, taking a diverse set of approaches. The next section will be devoted to a review of these hypotheses. In this review I do not attempt to be comprehensive and review *all* the literature about Zipf's law. There is simply too much for any person to cover: there exists a vast body of literature on Zipf's law and related (power) laws from fields as diverse as computer science, mathematics, physics, linguistics, music studies and veterinarian studies. What I will attempt to do is highlight the main themes in the discussion with respect to my starting point, linguistics.

2.2.2 Hypotheses for Zipf's law

Explanations for Zipf's law have been put forward on two levels. Traditionally, Zipf's law is seen as caused by forces (whatever they may be) on the textual level. Zipf's law in these hypotheses is thus directly related to the sequence of words that form the text, or, more often, to the sequence of letters and spaces that form the words of the text. But recently, more and more hypotheses have been put forward that explain Zipf´s law as due to our mental capacities, usually the organization or functioning of our mental lexicon. According to these theories, the lexicon is organized such that some properties of it follow a power law. Randomly sampling words from such a lexicon then automatically results in a text that conforms to

Zipf's law. Although intuitively appealing, convincing evidence that this is so is unfortunately still lacking.[3]

In what follows I will present an overview of the most influential hypotheses that have been put forward to explain the existence of Zipf's law.

The Principle of Least Effort

G.K. Zipf's own explanation for Zipf's law was that it is due to what he called the *Principle of Least Effort* (1949). He defined this principle for a range of fields. For language, it was formulated as a trade-off between hearer needs and speaker needs (whereby speaker and hearer should not only be understood as two different individuals, but also as alternating roles within each individual). For a speaker, the required effort is lowest when he or she can communicate any message with just one word. Zipf calls this the *force of unification* due to *speaker's economy*: "For having a single-word vocabulary the speaker would be spared the effort that is necessary to acquire and maintain a large vocabulary and to select particular words with particular meanings from his vocabulary" (Zipf 1949, p. 20). But this would be the worst-case scenario for the listener. For him or her, minimal effort is achieved when every meaning is conveyed through a different word. This is what Zipf calls the *force of diversification*: "This one-to-one correspondence between different words and different meanings, which represents the *auditor's economy*, would save effort for the auditor in his attempt to determine the particular meaning to which a given spoken word referred" (p. 21). This in turn would be the scenario that would maximize speaker effort. The two opposing forces that result from speaker's and auditor's economy shape the diversity of the vocabulary and create a situation in which about half of the speech tokens consist of very frequent types, while the other half are low frequency types. The result is a frequency distribution that conforms to a specific power law, later labelled Zipf's law.

This explanation is intuitively appealing, but Zipf failed to provide any rigorous mathematical proof for his theory. Rapoport (1982) for instance says that it is 'stated in vague, connotation-ridden language, precluding rigorous deduction' (p. 3). Lees (1959) in his review of Apostel, Mandelbrot & Morf (1957) uses even stronger words and says: "I shall not review Zipf's own explanation for his curves, since one

[3] Maybe mostly due to the fact that it would be very hard to provide such evidence. What could it possibly consist of? One possibility is to find mental lexicons that do not conform to Zipf's law. These might be found in certain patient populations, possibly semantic dementia.

can hardly make any sense of his so-called Principle of Least Effort” (p. 275). This imprecision and vagueness quickly gave rise to alternative theories, of which intermittent silence and Simon’s model are the most well-known.

Intermittent silence

One of the most influential alternative hypotheses for Zipf's law is that it is due to the random placement of spaces in a text.[4] It is in fact a group of explanations, known as *intermittent silence*. Two explanations stand out and are amongst the most well-known explanations for Zipf’s law. These are the explanations by B.B. Mandelbrot and by G.A. Miller.

B.B. Mandelbrot (1953)’s hypothesis is conceptually parallel to Shannon (1948)’s information theory about the transmission of information over a noisy channel, but more focused on the description of natural languages. It is based on minimization of the average cost per unit of information in a text, an explanation that is reasonably close to Zipf’s own explanation in terms of minimum effort. In this model, messages are generated one word at a time. During this process, the system maximizes the transition of information, while simultaneously minimizing the cost of transmission. The communicative cost of words is defined in terms of the letters that spell those words, and the spaces that separate them. All letters are considered equally costly. It is thus most cost efficient to use all the possible one-letter words most frequently, then to use all two-letter words, etc. In other words, the cost of words is proportional to their length, so the number of words of a given length in a language is also proportional to that length. A result of this model is that the message is basically generated as a random sequence of letters, separated into words by spaces. Mandelbrot showed that Zipf’s law follows as a first approximation from the minimization of communication cost when formulated in this way (Mandelbrot, 1954; see also Baroni, 2008; Manin, 2008; Miller, Newman & Friedman, 1958; and Wyllys, 1981).

Mandelbrot acknowledges that natural language is not the optimal coding system described above: it is not always the case that short words are most frequent. He therefore argues that letters (or phonemes) are not the proper units in which to

[4] These hypotheses are formulated in terms of written language, but are not restricted to that. Generally (and usually implicitly), it is assumed that we write the way we talk. Written language is in fact written down spoken language, and the two are therefore the same. Mandelbrot points out that the actual level on which Zipf's law works might be the phoneme or morpheme level, rather than the level of the characters in written text (Mandelbrot 1954, p. 217).

measure word length. According to Mandelbrot "one must therefore make the weaker assumption that the structure of speech as a sequence of words is influenced by some other coding, higher up in the receiving brain, considered as an optimal terminal information processing machine" (Mandelbrot, 1954, p. 217). The elements from this optimal coding are then imperfectly reflected in our letters and phonemes, while on average it is still the case that short words are much more probable than long words despite the many exceptions (Miller & Newman, 1958).

A conceptually simpler explanation for Zipf's law that arrives at the exact same outcome was put forward by George A. Miller (1957). It is so simple that it should be treated as a null hypothesis concerning how language may act in the absence of other forces (Piantadosi, 2014; Ferrer i Cancho & Solé, 2002). Just like Mandelbrot, Miller proposes a model that results in a string of letters separated by randomly placed spaces (he himself also acknowledges the identical output: Miller, 1957; Miller, Newman & Friedman, 1958). The difference between the two models is that in Miller's model, there is no need for optimization. His explanation became one of the most well-known and influential explanations for Zipf's law.

The argument runs as follows. G. A. Miller claims that there are no deep reasons for Zipf's law, but that a hypothetical monkey behind a typewriter could produce it just as well. Imagine a monkey that hits the keys of a typewriter at random, subject to the following constraints: (1) he must hit the spacebar with a probability of $p(*)$, and all the other keys with a probability of $p(\mathrm{L}) = 1 - p(*)$, and (2) he must never hit the spacebar twice in a row. In his output we will then find strings of i letters in a row, separated by spaces. The probability of a word of length i will decrease exponentially as i increases: if there are x words with one letter then there are x^2 words with two letters, x^3 words with three letters, etc. In other words: shorter words will be much more likely to occur than longer words. Words of the same length are equally probable. If we now look at the rank-frequency distribution of the monkey's output, we will find that it follows a power law. Miller claimed, therefore, that it is not surprising that our language output follows Zipf's law, since monkeys typing at random manage to do it as well as we do. According to him, Zipf's law could thus be derived from simple assumptions, without the need to appeal to least effort, least cost, or optimization principles (Miller, 1957).

Miller and Newman (1958) acknowledge the problem that this model gives a step-wise rather than a power law function. A solution for this is provided by Li (1992). He repeats[5] Miller (1957)'s argument that random text generators output a Zipf's-

[5] However, Li never mentions Miller (1957), where mathematical proof for the same point was already provided. He only mentions Miller's much more general Introduction to G.K.

law-like distribution of word frequencies by mathematically proving that this is the case, resulting in an exponent of $\alpha \approx 1$. He adds that the steps can easily be removed by introducing bias among different symbols, such that different symbols have different probabilities to appear in the sequence.

Miller and Mandelbrot both showed that the random placement of spaces leads to a power-law distribution of 'words'. The problems with these theories that have been pointed out in the literature broadly fall into two categories: (1) the distribution that follows from random models is *not* the same as real frequency distributions, and (2) philosophical, fundamental considerations question the plausibility of these models.

With respect to distributional problems, the frequency distribution that follows from random typing models in fact does not properly approximate that of natural language when statistically rigorous tests are used (Ferrer i Cancho, 2005a; Ferrer i Cancho & Elvevåg, 2010; Ferrer i Cancho & Solé, 2002). The slope of random texts is steeper, which means that it falls above the distribution of real texts for high frequencies and below it for low frequencies. Not a trace of steps is visible for the higher frequencies of real texts, while random texts clearly show a stepwise distribution (because words of equal length are equally probable). The typical slightly curved distribution of real texts is not seen for random texts. These differences are especially clear when large corpora are investigated (Tripp & Feitelson, 2007; Montemurro, 2001).

A second problem is that the random typing accounts predict that the number of word types of a given length should grow exponentially with word length, but this exponential distribution is not found for natural language: in fact, it looks more like a normal distribution with a long right tail (Manin, 2009). A related problem is that this model predicts a strong relation between length and frequency: short words have high frequencies of occurrence; long words have low frequencies. On average length does decrease with frequency (Miller, Newman and Friedman, 1958), but especially for shorter words large deviations for the predicted values are found: English uses too many three- and four-letter words and too few one-letter words, Russian has a bimodal distribution with peaks for one and seven letters (Oettinger, 1954). Also, it is predicted that words of equal length are equally probable, a prediction that is not met (Howes, 1968). And finally, a striking observation is that almost all of the very frequent words are grammatical words, not lexical words. For content words separately, only a slight preference for shorter words was found. If

Zipf's *The Psycho-Biology of Language: An Introduction to Dynamic Philology* (MIT Press, 1965).

the intermittent silence model was right, then it would be expected that the group of lexical words separately holds the same properties as the system as a whole. Since this is not the case, the intermittent silence model must at best be considered incomplete (Miller, Newman & Friedman, 1958). If random typing accounts do not correctly predict the length and frequency distributions of natural language then it must be the case that it is something specific to natural language that causes Zipf's law.

With respect to the more fundamental problems, it has frequently been claimed that models of random text generation do not capture the real process and therefore are poor scientific theories (e.g. Meyer, 2002; Howes, 1968; Piantadosi, 2014). The most obvious problem is that a real text is not made up from random letters. There are many theoretically possible combinations of letters that do not occur in natural language (Newman, 2005). It is unclear how any form of random typing could underlie or result in the meaningful string of words that humans usually utter. People use and process words in their entirety, not letter by letter. We do not distribute our word boundaries randomly over the text. On a textual level, our choice of words is bound by all sorts of syntactic requirements. As Ferrer i Cancho (2006, p. 132) puts it: "(…) [S]yntax is responsible to a great extent for the existence of correlations between words within real word sequences (Ferrer i Cancho & Elvevåg, 2005). Thus, it is striking that those who have largely defended syntax as the crux of human language (Hauser et al., 2002) argue that intermittent silence can explain Zipf's law in real human language." Also, we use words to convey meaning, something that is completely ignored by the random typing models (Piantadosi, 2014). These models can therefore never be a real explanation for Zipf's law. Clearly, a deeper explanation than random typing is required.

Preferential attachment

One of those deeper explanations came from H.A. Simon, a contemporary of Miller and Mandelbrot. His hypothesis, now known as *Simon's model*, was basically an extension of an older model from the 1920's by G.U. Yule, generally known as the *Yule process*. Yule was inspired by observations of the statistics of biological taxa. These biological taxa have a power-law distribution of sizes: the distribution of the number of species in a genus, family or other taxonomic group appear to follow a power law quite closely. Simon (1955, 1960) adapted this model to fit language. It is a rich-get-richer model that was originally formulated on a textual level, by assuming that a text grows by adding new words to the end of it whereby frequent words are preferred and re-used more often. However, it can be easily extended to concern networks, thus offering an explanation on the level of the mental lexicon.

Simon's model works as follows. Suppose we have a system composed of a collection of objects, in this case a text composed of words. New objects appear every once in a while, when new words are added to the text. Each object has some property *k* associated with it, which in this case is the number of times that word has been used before in that text. New objects have some initial value of *k* which will be denoted k_0. Usually, $k_0 = 1$: as soon as a new word appears it has been used once. The value of *k* increases when the word is used more often.

In between the appearance of one object and the next, *m* repetitions of previously used objects are added to the system. Words that are already frequent are more likely to be repeated. This is called a *rich-get-richer* process, also known as *cumulative advantage* (De Solla Price, 1976) or *preferential attachment* (Barabási & Albert, 1999). It can be mathematically shown that this process leads to a power law in the property *k*, in this case the number of words (for mathematical proof, see e.g. Simon, 1955, 1960; or Newman, 2005). In other words, some words end up being highly frequent, while most are used very infrequently. Simon emphasized that this hypothesis is compatible with the notion that people select words according to the current topic, rather than completely randomly (Simon, 1955). According to Simon, authors write by processes of association, i.e. sampling earlier segments of the word sequence, and by processes of imitation, i.e. sampling segments of word sequences from other works they have written, from works of other authors, and from sequences they have heard (Simon, 1955: p. 434). Association explains the higher frequency of words that are already used. New words can be added at any time, thus allowing for imitation.

This model is extended to cover lexical networks by reading 'nodes' where 'words' is written, which are connected to a lexical network with *k* number of connections. In between the appearance of new nodes, *m* new connections are added to the system, reflecting newly learnt meanings to already existing (previously added) nodes.

The crucial property of the system is that it is an open system to which new objects can be added over time, in other words, during the construction of the text. This also follows from the study of Simplified Chinese characters by Dahui, Menghui and Zengru (2005) that was already mentioned above. In Simplified Chinese, it is practically impossible to add new characters. Writers are thus restricted to the existing body of characters, which has a size of 8105 characters (but new phrases are perfectly possible). In comparison, the Second Edition of the 20-volume Oxford English Dictionary (1989) contains entries for 291.500 words (*Oxford English Dictionary, 2017*). This means that the maximum number of ranks any Simplified Chinese text can have is 8105. When this number is reached, a writer can only repeat the characters he used before. No such restriction exists for other languages, where one can always come up with new words. Dahui, Menghui and Zengru argue that it is this fixed number of characters that creates the deviation from Zipf's law that they found for Simplified Chinese: they found a curved shape that dropped much faster to 0 than that of English, in which mid-range frequency values were overrepresented. They conclude that it must be the (possibility of) lexical growth that causes Zipf’s law. However, they do not provide any mathematical proof, and thus we can only safely conclude that the possibility of growth seems to be a necessary condition for Zipf’s law, but not necessarily a sufficient condition.

A modified version of Simon’s model in texts was proposed by Montemurro and Zanette (Montemurro & Zanette, 2002; Zanette & Montemurro 2005). In Simon’s model, words are added at a constant rate α such that the vocabulary size at each time step t is $V_t = \alpha t$. Montemurro and Zanette add a parameter v to match the vocabulary growth of real texts better, such that $V_t = \alpha t^v$, with $0<v<1$. This parameter v is affected by two factors: author style and degree of inflection of the language. Texts in highly inflected languages (such as Latin or Finnish) contain many more word tokens than texts in words with fewer inflection (such as English), thus causing a larger vocabulary and a lower repeat rate. This is modelled by a larger value for v. Their other modification consists of introducing a word-dependent threshold for newly introduced words. The result of this threshold is that words that have been recently introduced have a slightly higher competition advantage and their reappearance in the text is favoured. They show that this model can closely match empirical data, which in this case consists of three novels in English, Spanish and Latin.

A general problem with these kinds of models is that we need to assume or deduce values for some parameters of the model, namely k_0, m and, optionally, c, some constant such that new connections appear with the proportion $k + c$ instead of simply k, to prevent that nodes with $k_0 = 0$ never get any new connections. In Montemurro and Zanette’s model we also need to assume values for v and for the parameter that determines the word-dependent threshold. With the right values for these parameters we can match the distribution of k that we encounter in real life. But there seems to be no good method to determine *a priori* what those values should be, thus falling to the risk of simple line fitting (Meyer, 2002). Another problem is that Simon’s model cannot explain exponents larger than unity, although they *are* found in empirical data, and the original model cannot explain the faster decay that is often found for low frequencies (Zanette & Montemurro, 2005).

Another objection is that, according to these models, it seems to be a game of chance which words are most frequent: some word might just by chance be used once more often than another word, but as a result of that the chance that it is used again rises. Here, too, it is unclear how processes in human language production could behave in this way, and still create a meaningful text (see also Piantadosi, 2014, for a similar argument). Frequent words are in fact likely to be used again, but it seems like syntax (for grammatical words) and topic (for lexical words) are responsible for this (Piantadosi, 2014).

Preferential attachment extended: Growth models

The problem of chance is elegantly solved by the model of semantic growth from Steyvers and Tenenbaum (2005) that was already briefly mentioned above. They make use of Simon's model to simulate the growth of semantic networks. They state that "under the semantic net view, meaning is inseparable from structure: The meaning of a concept is, at least in part, constituted by its connections to other concepts." Their model shows how such a meaningful structure can arise.

Steyvers and Tenenbaum take the following approach. First, they study networks based on three sources of real-world semantic knowledge. Networks can be either directed, meaning that node A can be connected to node B while the reverse is not necessarily true, or undirected, meaning that connections always go both ways. Steyvers and Tenenbaum built a total of four networks. Two of them, one directed and one undirected, were built based on word association norms from the free-association database constructed by Nelson, McEvory and Schreiber (1998). More than 5000 words served as cues in this project. A link between two words was established if one was ever used as an association to the other, and only words that were themselves cue words were included.

The third network was undirected and built based on Roget's Thesaurus. This thesaurus was constructed by Dr. Peter Mark Roget (1779-1869) and contains over 29.000 words classified into 1000 semantic categories. In the network, a link between two words is assumed if these words have at least one category in common. Classes are not included as nodes.

The fourth network was also undirected, and constructed based on WordNet, a large lexical database developed by Miller and colleagues (Miller, 1995; Fellbaum, 1998). In this database, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets) that each express a distinct concept. It contains more than 120.000 word forms and more than 99.000 word meanings. Word forms are connected to the corresponding word-meaning nodes (which can be more than one in the case of polysemous words), and to other word forms that are connected by relations such as hypernymy, meronymy or antonymy.

Steyvers and Tenenbaum find that these networks all possess a *small-world* structure, meaning that the distance between any two nodes (concepts) in the network is surprisingly small relative to the size of the network as a whole. This small-world structure seems to arise from a scale-free organization, the property that some nodes are highly connected while (most) others only have few connections. In a network, this means that some nodes function as hubs: they are very well connected and travelling from one node to another is highly facilitated by these hubs (much like a large train station with many connecting trains facilitates traveling from any one station to any other). This makes the average distance between any two nodes remarkably small. According to Steyvers and Tenenbaum, the fact that they find such similar organizational structures across semantic networks based on different kinds of data is likely to reflect, at least in part, some abstract feature of semantic organization: they argue that networks like theirs might directly form the basis for hypotheses about mental representations of semantic knowledge, or they may be an abstraction from some other representation (in which case the outcomes of their models constrain the possibilities of these more detailed models of semantic structure).

Based on their findings from these models, Steyvers and Tenenbaum propose a model of semantic growth. In this model, they start with a small, fully connected network (meaning that every node is connected to every other node) to which new nodes are added over time. New nodes connect to a subset of the nodes within an existing neighbourhood (defined as some node and all its connected nodes), with the probability of choosing a particular neighbourhood proportional to its size. The network can either be directed or undirected. If it is directed then the direction of each new node is chosen randomly. They find that this process naturally results in a scale-free, small-world structure.
Steyvers and Tenenbaum argue that the process they modelled can be viewed as a kind of semantic differentiation, in which new concepts correspond to more specific variations on existing concepts, and highly complex concepts (those with many connections) are more likely to be differentiated than simper ones. The process ensures a relation between the history of the network growth, and its ultimate pattern of connectivity. In other words, the point in time at which a node is added to the network is related to the number of connections it ends up with: words that were added early are most likely to end up with most connections.

The difference with the more traditional Simon's model of preferential attachment is that the model from Steyvers and Tenenbaum has more psycholinguistic reality to it. The result of the model is that early acquired words are most likely to end up with most connections. As Steyvers and Tenenbaum word it: "Most generally, our growth model predicts a correlation between the time at which a node first joins the network and the number of connections that it ultimately acquires. More precisely, at any given time, older nodes should possess more connections than younger nodes, and this effect should interact with variation in utility (e.g., word frequency) that influences the probability of connecting new nodes to particular existing nodes." They thus predict an interaction between the number of connections a word has and its utility, in other words, how available it is for usage. This means that early acquired words should be more available for usage, since these are on average the words with most connections. This could be a direct explanation for the well-known age of acquisition effect that was discussed in Section 2.1.2 above, the effect that words that are on average acquired early by children are responded to faster under all sorts of experimental conditions, such as lexical decision tasks or rapid naming experiments (e.g. Carroll & White, 1973; Turner, Valentine & Ellis, 1998). Having more connections means that activation that spreads through the network can reach a node more easily, thus raising its level of activation and thereby increasing its chance to be selected for production. It seems reasonable to assume that words that are produced more easily are also the words that are produced more frequently, because it requires less effort to use these words. If so, then this would explain the frequently found interaction between age of acquisition and word frequency.

Generally speaking, Simon's model and growth processes like the ones modelled by Steyvers and Tenenbaum provide a general and psychologically plausible mechanism that can explain a large variety of power law distributions, not only in language but also for example in author citations and links on the web (Newman, 2005).

New optimization models

Despite these other explanations, Zipf's original hypothesis was not forgotten. Present day technological advances have now made it possible to provide the proof for Zipf's Principle of Least Effort that was lacking from Zipf's original work. Crucial to this is the notion of *phase transition.* Phase transitions have been used to explain similar phenomena in multiple domains (e.g. Newman, 2005; Sornette, 2006), including language (Ferrer i Cancho, 2010).

Some network theory

Phase transitions are a phenomenon of networks. These explanations thus work on the level of the mental lexicon, which is considered to be a complex network. This network consists of *nodes*, which represent words, and *edges*, in other words the connections between them (or to other types of nodes). Two connected nodes are called *neighbours,* and a set of nodes formed by one node and all of its neighbours is called a *neighbourhood.* The number of edges connecting a node to its neighbourhood is called its *degree*.

Consider a growing network. If every node in this network is connected to only a very small number of other nodes (maybe to only one other node, if any), then new edges can easily be added at any time without having to add new nodes. The size of the network does not restrict growth of the number of edges. On the other hand, if nodes are all very highly connected (maybe to all other nodes), then new edges can only be added if the number of nodes grows. This means that the size of the network (in number of nodes) restricts growth of the number of edges.

Now, consider a sparsely connected network with a fixed number of nodes, to which new edges are being added. At some point during this process, the number of edges goes from the one extreme (being independent of network size) to the other extreme (being dependent of network size). This turning point is what is called the *critical point* (illustrated in Figure 2.3) (see also e.g. Farmer & Geanakoplos, 2008, for a more technical discussion of critical points and power laws). At or close to this point, microscopic changes to the system can result in macroscopic effects on a much larger scale. Crucially, only at this point, the probability for nodes to be connected to other nodes follows a power law. This specific probability distribution results in a scale-free distribution of connections: some nodes have very many connections to other nodes, while most nodes have very few connections to other nodes (in fact, about half of them have only one connection) (explanation adapted from Newman, 2005). This scale-free distribution of nodes is considered to be the relevant feature for Zipf's law to appear.

Phase transitions and Zipf's law

Why would the system remain in this critical stage? Why does the scale not tip to either of the two extremes? And how does all this relate to Zipf's law? The explanation for this is believed to lie in *self-organized criticality*, the property of some dynamical systems to arrange themselves such that they always remain at the critical point. It is this property that can cause the edges in networks to conform to

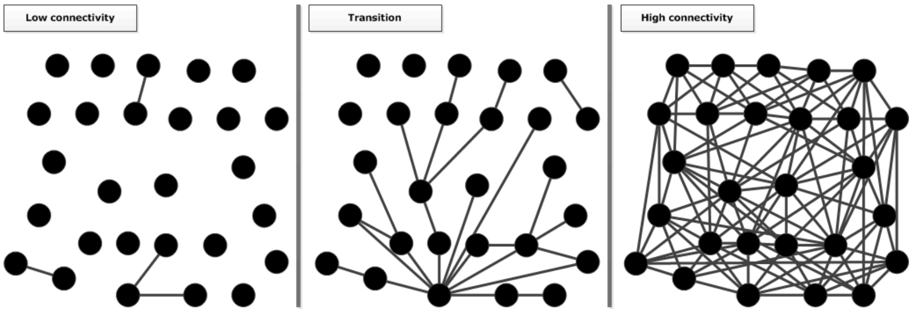


Figure 2.3. Illustration of phase transition

Zipf's law, as was shown by Ferrer i Cancho and Solé (2003). Their explanation for Zipf's law lies in the way the human mental lexicon evolved over time: they aim to model the process in which our ancestors went from some rudimentary referential signalling to the large lexicon we have today. The dynamical processes they describe thus do not take place in the mental lexicon when children acquire a language but took place over time as language developed in humans. The resulting lexicon is passed on from generation to generation.

Ferrer i Cancho and Solé propose a model of the lexicon in which the lexicon is composed of a set of signals (which roughly corresponds to words), a set of objects of reference (which roughly corresponds to meanings), and edges between these two kinds of nodes. An optimizing algorithm distributes the connections such that the entropy – a measure of complexity – is lowest for both speaker and hearer (which should again be seen as alternating roles, not two different persons). G.K. Zipf's concepts of speaker effort and hearer effort are further defined in terms of entropy, an information-theoretical measure for uncertainty: speaker entropy is lowest when one signal is used to refer to all objects of reference; hearer entropy is minimized when a different signal is used for every object of reference. These two opposing forces shape the distribution of edges over the lexicon. Ferrer i Cancho and Solé find that the lowest entropy for the system as a whole is obtained exactly at the critical point, the point where the degree distribution follows a power law: Zipf's law. Ferrer i Cancho and Solé thus claim that it is the constant need to balance hearer and speaker needs that causes Zipf's law to emerge in language, as did Zipf half a century earlier.

Ferrer i Cancho (2010) discusses this model a bit further. Here, he argues that the critical point at which Zipf's law is encountered is in fact a phase transition between a "no communication phase" and a "perfect communication phase". Language is then a system that is self-organizing itself between order and disorder (as are many other complex systems).

Piantadosi (2014), however, criticises the model from Ferrer i Cancho and Solé (2003). He argues that some choices in the design of the model are undesirable, such as the assumption that all objects of reference are equally likely. In later versions of the model (Ferrer i Cancho, 2005b) this is indeed no longer assumed: the probability of each object of reference (roughly, meaning) is now proportional to the number of signals (roughly, words) that are associated with it. [6] Ferrer i Cancho (2006) states:

[6] Although it could just as well be the other way round, with the number of signals being proportional to the probability of a meaning (A. Dimitriadis, personal communication).

"The two branches of models are very interesting from the philosopher's perspective in that one assumes that the frequency of what we talk about is dictated by the 'outside' world while the other leaves the frequency to the internal organization of the communication system itself. Tentatively, the first branch may seem more reasonable, but in fact, communication in human language is often detached from the here and now [...]." But according to Piantadosi, it is generally unknown what the real distribution should be, and thus to what extent these assumptions are correct. Another criticism is that there is no reason to assume that speakers' difficulty is proportional to the entropy over signals. Piantadosi argues that it should be proportional to the entropy over signals *conditioned on a meaning* since this captures the uncertainty for the psychological system.[7]

A more fundamental problem that Piantadosi points out is that models like the one by Ferrer i Cancho and Solé are based on communicative optimization. But Piantadosi (2014) argues that we also find Zipf's law in domains where this mapping is highly constrained by the natural world, meaning that little optimization can take place (e.g. names of months, planets or chemical elements, see Figure 2.4), and for number words (see Figure 2.6), where it is hard to imagine what such optimization would mean. According to Piantadosi, this does not mean that there is no way for a theory of optimization of communication to account for all the facts, but rather that such a theory has not yet been formalized.[8] However, it should be noted that data for Dutch show a somewhat different picture. In Figure 2.5 I plotted the frequency distributions of months, planets and chemical elements in SUBTLEX-NL, a database of Dutch subtitles. The names of months can still be argued to follow Zipf-Mandelbrot's law (although with a very high $\beta = 17$), but for planets and elements the distribution is notably different from the Zipf-Mandelbrot law. This might be due to outliers for the highest ranks, but with such small numbers of elements this is difficult to say. Thus, it can be questioned to what extent Zipf's law actually does hold for domains that are highly constrained by the natural world.

[7] Piantadosi also criticizes the specific value of the parameter λ, the parameter used to balance speaker and hearer entropy, at $\lambda = 0.4$. But this criticism is in my opinion not justified. This value was derived from the optimization algorithm, and not determined a priori. It has no special meaning, but the point is that it happens to be the case that this is the point at which the phase transition appears. The parameter is indeed very specific, but this is in my opinion no problem for the model as such.

[8] It should be kept in mind that 10 months (March and May were excluded because of interfering meanings), 9 planets (Pluto was still included) or even 116 elements (lead and iron were excluded) is not a whole lot to base such conclusions on, and visual inspection of Piantadosi's graphs casts some doubts on the fit of Zipf's law to these data.

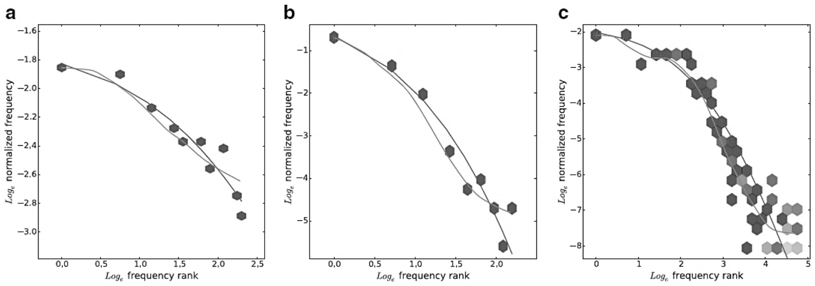


Figure 2.4. From Piantadosi, 2014. Frequency distribution in the American National Corpus for words whose scope of meaning has been highly constrained by the natural world: *a* months, *b* planets, and *c* elements. The dark grey line is the fit of Zipf-Mandelbrot's equation, the lighter grey line is a LOESS (locally weighted smoothing).

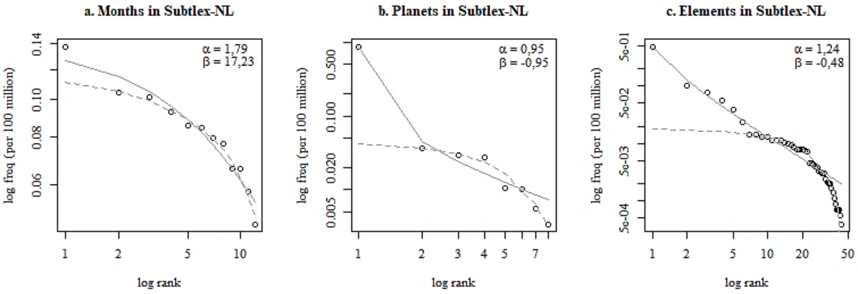


Figure 2.5. Frequency distribution in the Dutch subtitles corpus SUBTLEX-NL for words whose scope of meaning has been highly constrained by the natural world, following Piantadosi, 2014: *a* months, *b* planets, and *c* elements (only those 44 elements with non-zero frequency counts) ('boor', 'koper' and 'zink' were excluded because of their other meanings in Dutch). The solid line is the fit of Zipf-Mandelbrot's equation (notice the curve in the opposite direction for planets and elements), the dotted line is a LOESS.

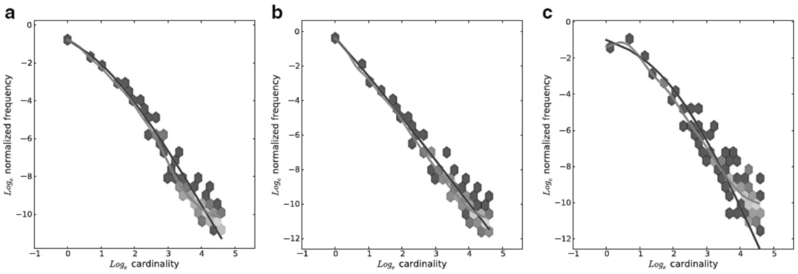


Figure 2.6. From Piantadosi, 2014. Power law frequencies for number words ("one,", "two," "three," etc.) in English (a), Russian (b), and Italian (c), using data from Google (Lin et al., 2012). Decades ("ten," "twenty," "thirty," etc.) were removed from this analysis due to unusually high frequency from their approximate usage. The dark grey line is the fit of Zipf-Mandelbrot's equation, the lighter grey line is a LOESS.

Corominas-Murtra and Solé (2010) use a different optimization model, conceptually more closely related to Simon’s model of preferential attachment. They do not look at the number of connections in a (theoretical) mental lexicon but focus on the textual level by modelling the actual string of words. They stress that language is best modelled by an open model, in which the system can encounter new possible outcomes at any time (unlike for example a dice rolling experiment, where the possible outcomes are known *a priori*). They assume a model in which the distribution of probabilities of the possible outcomes in the model maintains its basic statistical properties as it grows (so the most likely outcomes will remain most likely, also when new elements are added). Corominas-Murtra and Solé are especially interested in the case where there is some balance between ordering and disordering forces, such that the output of the system (in other words, the text) displays entropy values (the number of bits necessary to code any element in the text) proportional to the maximum entropy achievable for the system in equilibrium (in other words, the entropy value when disorder is maximized, at a point in time where the parameters that characterize the system are not changing). Put simply, this means that they are interested in the case where the order of the words in the text that the system produces is not completely random, but not completely predictable either. Rather, it occupies some place in between these two extremes. Their question is what the probability distribution looks like at this point of equilibrium.

Corominas-Murtra and Solé present two ways to solve their problem. Their first strategy is purely mathematical. They assume that the probability distribution follows a power law, and what they want to know is the value of the exponent. What follows from their calculations is that the only solution is that this exponent equals one, the traditional value of Zipf’s law.
Their second strategy is linguistically more realistic. They assume that the mechanisms responsible for the growth and stabilization of the system do not depend on the size of the output. In other words, the mechanisms responsible for Zipf's law in a short text will be the same as those responsible for Zipf's law in a long text, so studying what you have at any given point in time will allow you to determine the values you get when the text grows. This means that a partial observation of the system is enough to infer the probability distribution of the possible outcomes of the system as a whole. They call this the *scale invariance condition*: the entropy restriction (intermediate between maximal and minimal) works on all levels of observation, independent of scale. They mathematically show that Zipf’s law is the unique asymptotic solution to such a system. From this they

conclude that Zipf's law is the natural outcome of a growing system in an intermediate state between order and disorder.[9]

In conclusion, these hypotheses of Zipf's law as the result of a system in a critical phase between order and chaos are generally very elegant and intuitively appealing. However, the general problem is that a lot of assumptions have to be made about the precise workings of the system to get the models to work. It is unclear to what extent these assumptions are in line with the psycholinguistic reality.

Semantics

There is one other branch of explanations to Zipf's law, which takes a very different approach from the previously discussed hypotheses. The one feature that sets human language apart from the output of the language models that were discussed above is that it has meaning. Some have argued that it is the way in which meanings are organized that is crucial for Zipf's law. As Guiraud (1971, p. 153) formulates it: "Zipf's distribution would be PRODUCED by the structure of the signified [the meaning] but would be REFLECTED in that of the signifier [the word]."

Guiraud (1968, 1971) was the first to propose a semantic hypothesis for Zipf's law. The system that he envisioned works as follows. Word concepts are the result of the combinations of a system of discrete *semic units* or *semes,* which are the smallest elements of meaning. These semes are organized in binary pairs, such as for example 'animate′ and ′inanimate', ′actor′ and ′process′, 'masculine' and 'feminine', etc. Each seme can be positive, negative or unmarked. The number of marked semes a word has (either as positive or as negative) determines its frequency, such that the probability of a word diminishes with its number of semes. If all semes are equally likely and the probability of a seme is equal to *p*, then that of a word of 1, 2, 3 *n* semes is equal to p, p^2, p^3, p^n: a system that leads to Zipf's law in words (but: with a stepwise distribution similar to that of Millers random typing account[10]). Guiraud's system is elegant, but not realistic. According to Guiraud, *dog* has to be much less frequent than *mammal* because of the fact that it is a strict subset, and thus narrower. But the opposite is true (frequencies in SUBTLEX-NL:

[9] This explanation for Zipf's law works not only for language, but can be extended to Zipf's law in other systems.

[10] Thanks to A. Dimitriadis for this observation.

hond (dog): 168,65 million; zoogdier (mammal): 0,96 million) (example from Manin, 2008, p. 1082).

D.Y. Manin (2008) points out another problem. One of the assumptions of the model is that all semes can be freely combined. But this cannot always be the case. If a word is marked for *process* in the *actor/process* distinction, then it is a verb. This means that *animate/inanimate* has to be unmarked: at least in western European languages there are no animate or inanimate verbs (by which Manin means that there is no verb that differs from for instance *laugh* or *worry* only in that it is inanimate).

Manin (2008) proposes a different semantic explanation. He assumes some semantic space *S*. This semantic space is “the set of all meanings”, without posing any significant assumptions about the nature of meaning. Word meaning is defined as words *w* taking up some volume in this space, thereby forming a subset of *S*. Words can thus be more generic or more specific by taking up more or less space in *S*, respectively. He assumes that the more generic the meaning, the more frequent the word, and vice versa, the more specific the meaning, the less frequent the word. In other words, speakers select the most imprecise word that still allows disambiguation in the given context. Less precise words are thus more frequent, and thus more accessible for both the speaker and the listener. This organization leads to a hierarchical classification of meanings over S, which gives rise to Zipf's law: if word 1 covers the whole of *S*, words 2 and 3 cover one-half of *S* each, words 4 through 7 cover one-quarter of *S* each, etc. then this automatically leads to a power-law mapping of meanings over *S* (Figure 2.7). As explained in Section 2.2.1, this power law behaviour is what characterizes Zipf's law. Of course, language does not follow this mapping exactly, but the mapping is close enough to result in Zipf's law.

The question is, then, why the mapping of words follows a Zipfian distribution, and not any other distribution. According to Manin, this is because of two processes: the expansion of meanings and the avoidance of close synonymy. Word meanings change as language evolves, which can be due to three more or less distinct processes: extension, formation, and appearance of senses. Manin argues that meanings tend to increase in scope (extension), unless they collide with other meanings of a similar scope. In this case, one of the words typically takes on a more specific meaning, thereby avoiding synonymy. An example is the interaction between *bread*, which initially meant ‘crumb, morsel’, and the word *loaf*, in Old-English *half*, which meant ‘bread’. When the meaning of *bread* extended to its current meaning of ‘bread in all its forms’, the meaning of *loaf* narrowed to something like ‘bread that is shaped and baked in a single peace’. Meanings of significantly different scope (where one has a more generic meaning than the other)

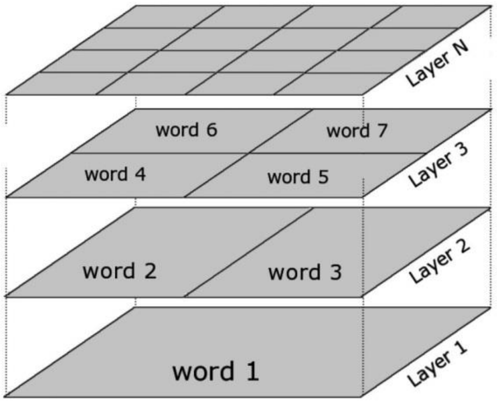


Figure 2.7. Example of a hierarchical organization of semantic space (Manin 2008: p. 1084)

do not interact. According to Manin, these processes result in a semantic space being covered almost without gaps and overlaps.

One weakness of this model is already pointed out by Manin himself: this model is based upon a rather vague theory of meaning. Many of the concepts on which this hypothesis is based are defined so vaguely that it is hard to see how this model can either be tested or falsified. Richie, Kaufmann and Tabor (2014), however, found a way to operationalize these concepts. “In their approach, a word type covers a region in space by virtue of the distribution of its *tokens* in that space,” as Richie (2016) puts it. To achieve this, they first count word-word co-occurrences in a corpus, in this case the British National Corpus (Table 2.1). This results in vectors that represent the meaning of each word type (Figure 2.8). They then pass through the corpus again, and for each token of each word type, they sum the type vectors of the other words in its context (defined as a moving 15-word window). This results in a vectorial representation of the meaning of that token. Each word type is thus represented by a cloud of tokens in semantic space. Finally, for each type, they compute the convex hull (the smallest set containing all items) around the token cloud (Figure 2.9). The volume of the hull measures semantic breadth; the facets of the hull localize it to a region of semantic space. They applied this method to 1000 tokens for each of the 1000 most frequent content words in the corpus. What they found was that word volumes indeed correlated with word frequency with more voluminous words being more frequent, as predicted by Manin’s hypothesis. Manin also predicted that volumes follow a power law distribution, namely Zipf's law. Evidence for this was ambiguous: they found that the 239th words with the most volume indeed fit a power law (after this, power law behavior ceased to exist), but a power law distribution with cut-off was no more or less likely than a log-normal

Table 2.1. Example of word-word co-occurrences

	Star	Drum
Planet	8	1
Comet	4	0
Britney	5	10

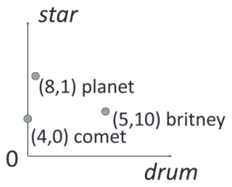


Figure 2.8. Example of vectors representing word meaning

distribution or an exponential distribution (see Figure 2.10 for an illustration of these different curves). The authors do not specify if the organization of these volumes follows Manin's predictions. This study thus provides evidence for some of Manin's predictions, but not for all, and not unambiguously so.

Another problem is the extent to which Manin's results actually follow Zipf's law. This claim was tested by Lestrade (2017), who finds that the distribution following

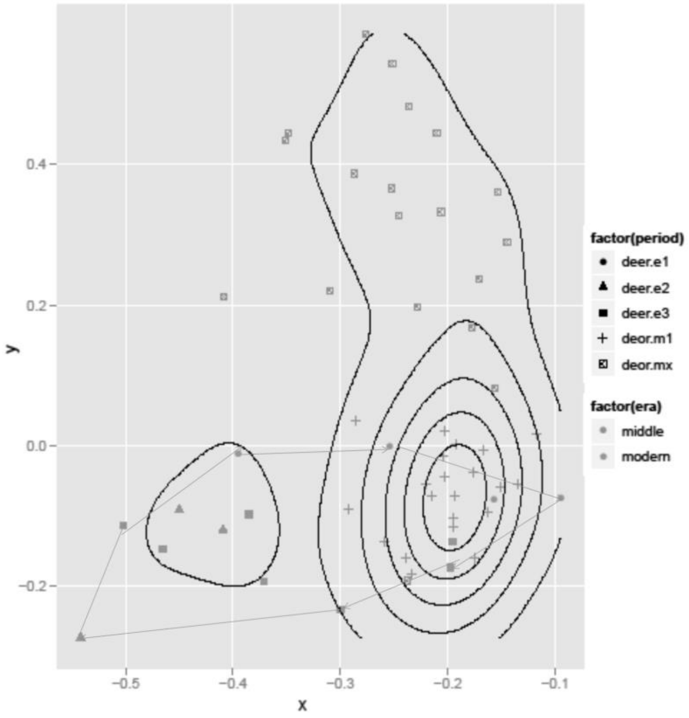


Figure 2.9. Semantic space of the word *deer* with token clouds in Early English (Middle Ages) (right cloud) and Modern English (left cloud). Each data point refers to the word in a different context.

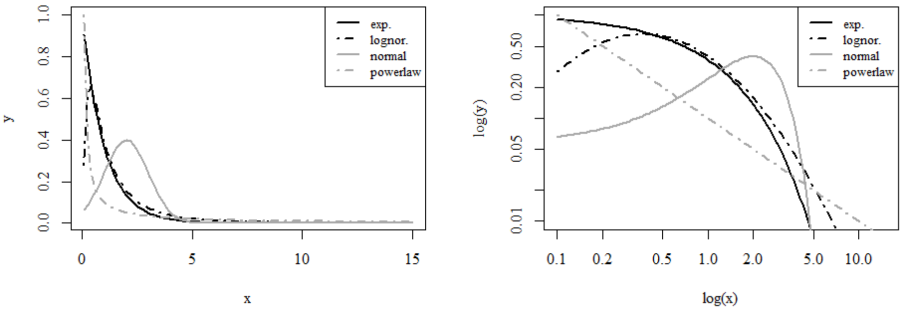


Figure 2.10. Illustration of the typical curves of four different distributions on normal and doubly logarithmic scales: an exponential distribution (solid black line), described by $f(x) = ab^x$ (here, a = 0 and b = 1); a lognormal distribution (dashed black line), the probability distribution of a random variable whose logarithm is normally distributed – in other words, if the random variable x is lognormally distributed, then $y = \ln(x)$ has a normal distribution (here, mean = 0 and sd = 1 on the log-scale); a normal distribution (solid grey line, mean = 2 and sd = 1); and a power law distribution (dashed grey line), described by $f(x) = ax^r$ (here, a = 0,1 and r = -1).

from Manin’s hypothesis actually curves the wrong way and has a much shallower slope than the one usually found for natural language. The graphs in Manin’s paper obscure this fact by using non-equivalent axes.

Finally, as discussed above, Piantadosi (2014) provides evidence that in American English also the words used for labelling the months, the planets and the elements follow a Zipfian distribution. These meanings are highly constrained by the natural world, so it is unlikely that their usage is in any way constrained by any sort of optimizing pressure. It is unclear how these findings could be explained by the semantic theories.[11]

Recently, yet another explanation was proposed for Zipf's law, which claims that it is the combination of syntax and semantics that causes Zipf's law to appear. This explanation, by S. Lestrade (2017), is not a modern formulation of one of the earlier hypotheses, but truly takes a different approach.
Irrespective of the language, words are divided into word classes or parts of speech (POS). The different parts of speech are used with comparable frequencies, but the size of these classes differ tremendously. This automatically results in much higher

[11] But then again, this would be difficult for *any* theory of Zipf’s law. It is likely that for example months are not equally likely to occur in our discourse. We are for example more likely to talk about April than February, because of April Fools’ Day. It is these kinds of factors that weigh in when considering these kinds of theories: it is not necessarily about how things *occur* in the natural world, but much more about how we *structure* our world.

frequencies for the words from the smaller classes (e.g. determiners). But this fact alone is not enough to reproduce Zipf's law, as Lestrade shows: the resulting distribution contains discrete frequency bands that do not line up (Figure 2.11). Considering semantics, a word only becomes frequent if it is specific enough to single out its referent in context, but general enough to be applied to different referents. Using depth of embedding in WordNet, Lestrade shows that the most frequent words in the Brown corpus are indeed those words that are neither too general, nor too specific. To test if this feature alone is enough to reproduce Zipf's law, Lestrade creates an abstract lexicon in which words are specified for a number of meaning dimensions or vectors, features that seem to be somewhat comparable to Giraud's theory of semes.[12] These meaning dimensions can be specified as present or absent, or be left unspecified, and should be understood as representations of activation in a neural network model of the brain. The usage of words is then modelled by randomly generating contexts, containing a target object and a set of five distractors that are all fully specified for all meaning dimensions. A word is then selected from the lexicon that suffices to single out the target object. The result is – as expected – that words that are neither too specific nor too general become most frequent (Figure 2.12). The distribution departs from the ideal Zipfian distribution, though: the distribution is much more curved, and the slope is shallower than that of Zipf's law.

Syntax alone (more specifically, parts of speech) or semantics alone thus do not result in Zipf's law. The combination of both, however, does result in Zipf's law (Figure 2.13). Discrete frequency bands are no longer visible, and the curvature of the graph has disappeared. Unfortunately, Lestrade does not provide any parameter values nor does he provide goodness of fit measures of his models. This renders it difficult to validate the outcomes of his models. Nevertheless, the approach is promising.

2.2.3 Summary

Five branches of explanations for Zipf's law were discussed: the classic explanation of the Principle of Least Effort, explanations based on intermittent silence, the more modern approaches of preferential attachment, new optimization models building onto Zipf's Principle of Least Effort, and the varied group of explanations involving semantics. The Principle of Least Effort was found to be too undefined to have full

[12] Although Lestrade himself does not mention Giraud in his discussion of his model, while he does so in one of the files in the supporting information to his paper.

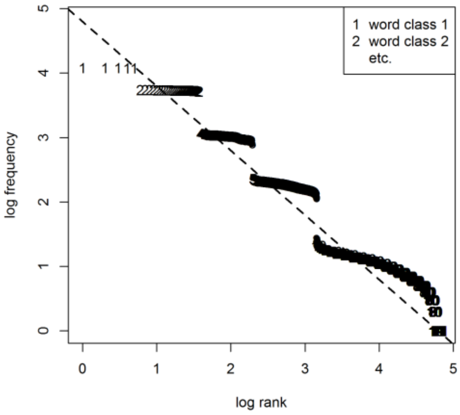


Figure 2.11. Word frequencies modelled solely based on POS (Fig. 2 in Lestrade, 2017)

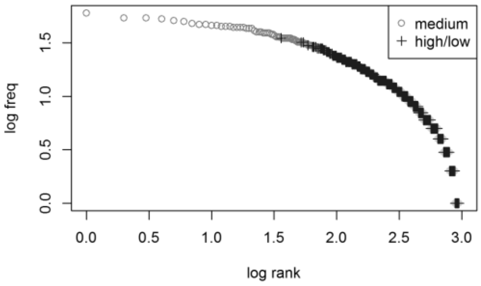


Figure 2.12. Word frequencies modelled solely based on semantics
(upper part of Fig. 4 in Lestrade, 2017)

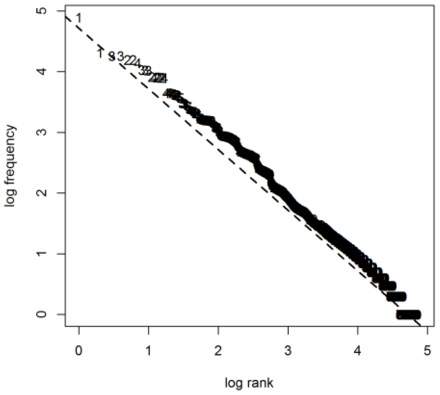


Figure 2.13. Word frequencies modelled based on the combination of POS and semantics
(Fig. 6 in Lestrade, 2017)

explanatory adequacy. New optimization models fill this gap, but they seem to depend heavily on specific settings of the computational models involved. Intermittent silence models can be used as a null-hypothesis concerning Zipf's law in the absence of any other forces, but they are unrealistic as true models of language production.

Except for Lestrade’s explanation regarding the combination of syntax and semantics, semantic models are not yet able to reproduce Zipf's law as it is found in human language. They can therefore not (yet) be relied on to explain the existence of Zipf's law. Lestrade’s model does seem to reproduce Zipf's law but he fails to provide any parameter values or measures for goodness of fit. Most promising and close to psychological (and possibly, neurological) reality is the group of explanations in terms of preferential attachment during network growth. They are not only capable of explaining Zipf's law in natural language but can also explain similar phenomena in other fields in which Zipf's law is found.

2.3 Discussion

Above, I concluded that the hypotheses explaining Zipf’s law through preferential attachment were most plausible. These theories place Zipf’s law in the organization of networks of semantic or possibly lexical knowledge. Semantic knowledge in Levelt’s model is accessed at the stage of the conceptualizer, while different parts of the lexicon are accessed at the stages of lexical selection, morphological encoding and phonological encoding and syllabification.

Semantic knowledge is generally thought to be intact in people suffering from aphasia. This means that if Zipf’s law originates at the conceptualizer, it should be unaltered in speech from people with word form anomia but might be compromised in speech from people with a form of semantic anomia. On the other hand, if Zipf’s law originates in the lexicon, it should be unaltered in speech from people with semantic anomia but might be compromised in speech from people with word form anomia. More research is necessary to know if these hypotheses should be confirmed or rejected. This study aims to be a first step in this.

The preferential attachment hypotheses reserve a large role for the age at which a word is acquired. Early acquired words shape the form of the network and are more likely to be more deeply embedded in the structure. This would also render them more resilient to damage to the network. If this is true, then it is expected that people with aphasia display more frequent usage of early acquired words as opposed to late acquired words. This question, too, will be addressed in the current study.

2.4 Scope of this study

In this dissertation, I only focus on non-progressive forms of aphasia caused by acquired brain damage such as a stroke. In line with Van Ewijk (2013), I do not classify patients according to the classic taxonomy but will adopt a broad distinction between fluent and non-fluent aphasia. My focus is on non-fluent aphasia, which includes both what Laine and Martin (2006) call semantic aphasia and what they call word form aphasia. I acknowledge that this includes a large variety of disorders, different for each individual patient. I will show, however, that this broad criterion is specific enough to leave me room to say something about the workings of Zipf's law, and to some extent the general organization of the mental lexicon. Presentation of individual patient data, where possible, will allow the reader to verify for him- or herself whether variation between patients allows for generalizations.

3 Explorations of Zipf’s law

3.1 Introduction

Zipf’s law has been studied extensively in large texts (e.g. Ha, Stewart, Hanna, & Smith, 2006; Hatzigeorgiu, Mikros, & Carayannis, 2001; Popescu, Altmann, & Köhler, 2010; Tuzzi, Popescu & Altmann, 2009). This law is considered a true linguistic universal in the sense that it is found in all texts in all languages (e.g. Montemurro & Zanette, 2011; Jäger & Van Rooij, 2007). Nevertheless, there are still many unanswered questions with respect to the exact behaviour of Zipf's law.

The original formulation of Zipf's law and of Zipf’s β-law are

1) $f(w) \approx \frac{C}{r(w)^{\alpha}}$ and 2) $n_f \approx C \cdot f^{-\beta}$

where in (1) $f(w)$ is the frequency of the word with rank $r(w)$ if words are ordered by decreasing frequency and C is a constant that is mainly dependent on text size and where in (2) the number of words n with frequency f, in other words, the size of the frequency class, is determined by parameter β and a text size dependent constant C. The parameters α and β were assumed to be constants with the values $\alpha \approx 1$ and Zipf’s $\beta \approx 2$ (Zipf, 1949, see also Chapter 2, Section 2.2.1 and the box on the next page). These traditional values of the exponent have been confirmed for many different kinds of texts in many different languages. Baroni (2008), for example, shows $\alpha \approx 1$ for the Brown Corpus of American English (Kucera & Francis, 1967), the written part of the British National Corpus, the novel *The War of the Worlds* by H.G. Wells (1898), the *La Repubblica* corpus (an Italian newspaper corpus; Baroni, Bernardini, Comastri, Piccioni, Volpi, Aston & Mazzoleni, 2004), and a Japanese web-page corpus (Baroni & Ueyama, 2006). Hatzigeorgiu, Mikros and Carayannis (2001) show that $\alpha \approx 1$ in the Hellenic National Corpus, a corpus of Modern Greek. Balasubrahmanyan and Naranan (2002) show that $\beta \approx 2$ in American Newspaper English, four plays in Latin by Plautus, a short story in Russian by Pushkin, colloquial Chinese, *Ulysses* by James Joyce and a sample of English texts in Dewey.

Terminology on Zipf's law

The terminology used in this chapter is the same as elsewhere in this dissertation, but is repeated here for convenience and in an attempt to limit confusion to a minimum (see also Chapter 2, section 2.2.1).
The term ‘Zipf's law’ is used as an overarching term for two formulas:

Zipf-Mandelbrot’s law:

$$1)\quad f(w) = \frac{C}{(r(w)+\beta)^{\alpha}}$$

where the frequency *f* of word *w* is determined by its rank *r* when all words are ordered from most to least frequent, the parameters α and β, and a text size dependent constant *C*, and

Zipf’s β-law:

$$2)\quad n_f = C \cdot f^{-\beta}$$

where the number of words *n* with frequency *f*, in other words, the size of the frequency class, is determined by parameter β and a text size dependent constant *C*.

Both Zipf-Mandelbrot’s law and Zipf’s β-law are formulations of the same phenomenon, and the unspecific term ‘Zipf's law’ has throughout the literature been used to describe both. To avoid confusion, when either of the two laws is discussed specifically it will be called by its full name.
Another potential point of confusion is the parameter β, because both laws have one. Therefore, the β from Zipf-Mandelbrot’s law will be dubbed ZM-β, while the β from Zipf’s β-law will be dubbed Z-β.

The only clear exception seems to be formed by the complete works of Shakespeare, with Zipf’s β = 1,6.[13]

Nowadays, the exponent of Zipf's law is generally considered to be a parameter instead of a constant and fitted to empirical data (e.g. Zanette & Montemurro, 2005; Baixeries, Elvevåg & Ferrer i Cancho, 2013; see also Chapter 2, Section 2.2.1). In comparing texts with the same number of tokens, a larger value of α then indicates a steeper slope of Zipf's law, meaning that frequencies drop faster when rank increases. This is an indication of a smaller vocabulary. Smaller values of α, on the other hand, indicate a shallower slope of Zipf's law, meaning that frequencies drop at a slower pace which thus indicates a larger vocabulary. The opposite holds for

[13] Balasubrahmanyan and Naranan suggest that this result perhaps reflects Shakespeare’s uniqueness in the literary world, although it is unclear what this would mean exactly.

Z-β. In this formulation of Zipf's law, frequency classes are used. The largest class is always that of words with frequency 1, the hapax legomena. A steeper slope of Z-β means that the number of tokens per frequency class drops faster, meaning that the smallest frequency classes are proportionally larger. This thus indicates a larger vocabulary. A shallower slope of Z-β means that the number of tokens per frequency class drops at a slower rate. Higher frequency classes thus contain more tokens, indicating a smaller vocabulary.

One obvious source of variation of the exponent – although little systematically researched –is the content of the text under consideration. A text with many different topics is likely to have a larger vocabulary than a text with only a few different topics. This shows for instance from the speech of people with schizophrenia, studied by Piotrovskii, Pashkovskii and Piotrovskii (1994, discussed extensively below). Amongst other populations, they studied speech from people with fragmented schizophrenia, characterized by multiple topics and the absence of a consistent subject, and advanced schizophrenia, in which speech is characterized by obsessional topics and related words and word combinations. For the patients with fragmented schizophrenia, they find $0{,}65 < \alpha < 0{,}9$. For the patient with advanced schizophrenia, they find $\alpha = 1{,}5$.

Another important source of variation of the exponent is text size, as was shown by Baayen (2001). He compared the slope of Zipf's law in the first 13250 words of Carrol's *Alice in Wonderland* to that of the full 26505 words. In the first case, it is found that $\alpha = 1{,}119$; in the second case $\alpha = 1{,}205$. He shows that this increase is systematic when text size grows, following an exponentially decaying function. The slope thus gradually becomes steeper. This finding reflects the fact that when text size increases, fewer new word types are being introduced to the text: most words have already been used before.

A third source of variation of the exponent is the language under consideration. Gelbukh and Sidorov (2001) compare English and Russian: they calculated the exponent for as many as 39 literature texts from different genres for each language. Each text was at least 10,000 words long. For English, they find an average of $\alpha = 0{,}974$; for Russian they find $\alpha = 0{,}893$. The smaller value for Russian reflects the fact that Russian is a morphologically rich language while English is considered morphologically poor: Russian thus has more inflected forms, resulting in more different word types in any text and thus a shallower slope of Zipf's law. Ha, Steward, Hanna & Smith (2006) compare English, Spanish, Irish and Latin. Their data is taken from the North American News Text corpus (489 million words), the Spanish corpus NLPSR 1.0 (16 million words), a corpus of 17th and 18th century Irish from the Royal Irish Academy (7,122,537 words) and the Latin corpus from

Harvey, Devine and Smith (1994) (2,244,444 words) (they compared thus corpora of different sizes). For ranks < 5000, they find the traditional $\alpha \approx 1$ for all languages (unfortunately, they do not present more exact values). For the English and Spanish corpus, they find that the Zipfian curve bends down at about rank 5,000 to $\alpha \approx 2$.
A downward curvature for the higher ranks has been reported more often (e.g. Ferrer i Cancho & Solé, 2002, see also Chapter 2, Section 2.2.1). Harvey and colleagues suggest that the differences between strongly inflected languages (Latin and Irish) and largely analytic languages (Spanish and English) might be the reason for these results, although they do not specify what property of these languages exactly causes this two-step behaviour. Others argue that it is the difference between lexical and grammatical or open and closed class words that cause this difference in slope (e.g. Ferrer i Cancho & Solé, 2000; Popescu, Altmann & Köhler, 2010). Both classes of words individually have a different slope, which results in a curvature of the frequency distribution when combined. Yet another option is that the curvature is being caused by the difference between a kernel lexicon formed by a language dependent number of versatile words (ca. 5000-6000 for English) and the rest of the lexicon for specific communication (Ferrer i Cancho & Solé, 2000).

The suggestion that it is the degree of inflection that causes the differences in slope is tested by Bentz, Kiela, Hill & Butterly (2014, the version of Zipf's law applied is Zipf-Mandelbrots law). Bentz et al. compare parallel translations of the Book of Genesis in Old English and in Modern English. Old English makes use of synthetic inflectional marking whereas Modern English encodes the same information using analytic constructions. This means that Old English used a high number of bound morphemes, thus creating many low frequency words, whereas Modern English uses a smaller number of high frequency function words. They find significant differences between the two languages that are independent of total number of words, style of translation, orthography or contents. For Modern English Genesis they find α = 1,22 for the whole text and α = 1,18 when text size is cut to the length of the Old English text, for Old English Genesis they find α = 1,03. For a lemmatized version of the Modern English text (so without any grammatical marking) an even steeper slope of α = 1,29 is found. The biggest impact on the shape of the frequency distribution thus comes from its ‘inflectional state’, the way in which inflection is encoded in the language.
Additionally, Bentz et al report on the values of Mandelbrot’s β (dubbed here ZM-β). As was discussed in Chapter 2 (section 2.2.1, see also the box above), ZM-β was introduced by Mandelbrot to account for the typical curvature of Zipf's law for the highest ranks, for which the traditional Zipf's law is known to overestimate the frequencies. Usually, ZM-β takes on some small value when ZM-α is close to 1. It seldom exceeds the value of 10. For Modern English they find ZM-β = 5,29, for Old English they find ZM-β = 1,49 and for lemmatized Modern English they find

ZM-β = 6,16. A higher Z-β-value means that for higher frequencies the frequency distribution deviates more strongly from the slope predicted by the ZM-α values given above. In other words, the less inflection, the more pronounced the curvature of Zipf's law for the highest frequencies. According to Bentz and colleagues, this difference arises mainly due to high frequency words (which are always function words) being used even more frequently in Modern compared to Old English.

The usage of the exponent of Zipf's law as parameters fitted to data is most clear in cases where Zipf's law is used to study special populations. Piotrovskii, Pashkovskii and Piotrovskii (1994, already mentioned briefly above) studied speech and written texts from six people with schizophrenia. Two types of the disorder were identified: fragmented schizophrenia (five patients), in which speech is characterized by multiple topics and the absence of a consistent subject, and advanced schizophrenia (one patient), in which speech is characterized by obsessional topics and related words and word combinations. Piotrowski and Spivak (2007) report on what seems to be the same patients, but with different values. In addition, Piotrowski and Spivak analyse speech from four children with down syndrome and from a group of 36 women, observed 1-2 weeks before and 3-4 days after giving birth. This group was considered to be under "birth stress" who are supposed to have a "limited number of topics left […] to discuss in the final stage of pregnancy" (Piotrowski & Spivak, 2007, p. 551). Word frequencies were studied using Zipf-Mandelbrots law. They compared the patient data to literary texts, scientific technical prose and combat documents, all in Russian. All texts were obtained either spontaneously or through questionnaires or structured talks and could be both written or spoken. Unfortunately, sample sizes per person/text were highly diverse (going from 294 tokens for one of the Down-children to 2.000.388 tokens for the technical texts) and different in all cases, which troubles comparison of these values. This should be kept in mind when interpreting the values reported below.
For normal Russian spoken language, they find α = 1,12. For the patients with fragmented schizophrenia, they find 0,65 < α < 0,9. For the patient with advanced schizophrenia, they find α = 1,5. The shallower slope for the patients with fragmented schizophrenia reflects their larger vocabulary due to the multiple topics, while the steep slope of the patient with advanced schizophrenia reflects the small vocabulary due to topics of obsession. For the children with Down syndrome it is found that 1,01 < α < 1,24. Some of them thus display a normal value of α, while others have values higher than the typical α = 1 thus reflecting a somewhat smaller vocabulary. The women under 'birth stress' display the same α-value before and after giving birth, in both cases α = 1,10, a value very close to the value of the control group. Literary texts displayed exactly the traditional α = 1,00. For technical texts, it was found that α = 0,83; for military texts α = 1,4. This means that vocabulary is more varied for military texts than for technical texts.

Hernández-Fernández and Diéguez-Vide (2013) studied Zipf's β in people with Alzheimer's disease, taking in speech from 20 patients in two different stages of the disease and from 10 controls. They looked both at all data points of the distribution, and at Zipf's β while only including the centre of the distribution (see Chapter 2, Figure 2.2B for an example of a Zipf's β-law distribution), where the data points follow a straight line when plotted on double logarithmic scales (some small deviations are often found for the highest and smallest frequency classes, the words that are most and least frequent). They found a value of Z-β = 1,15 or Z-β = 1,07 (all data points and centre of the distribution, respectively) for patients with moderate cognitive decline (4225 tokens) and Z-β = 1,17 or Z-β = 1,10 for matched controls (4913 tokens); and Z-β = 1,29 or Z-β = 1,40 for patients with moderate/severe cognitive decline (1528 tokens) against Z-β = 1,31 or 1,14 for matched controls (1481 tokens). The difference in number of tokens renders a comparison between the two Alzheimer's groups cumbersome. For the whole distribution, the differences between the two groups of people with Alzheimer's disease and their respective control groups were not significant. For the centre of the distribution, the more severe group displayed significantly higher Z-β values. As discussed above, this steeper Z-β slope is an indication of a more diverse vocabulary in this group, contrary to what might be expected. The authors argue that it might have to do with the disintegration of syntax in these patients.

Baixeries, Elvevåg and Ferrer i Cancho (2013) studied the development of α (traditional Zipf's law) in a large number of speech fragments from the CHILDES database (MacWhinney, 2000). Speech from Dutch, English, German and Swedish children between 1-3 years old was examined, and from the other people in the recorded conversations. For each speaker, the first 500 types that were uttered were examined. They found that α tends to decrease over time in the target children, a finding that reflects their increased vocabulary. A decrease of α was also found for their mothers, but less pronounced. For the other speakers, a significant correlation was found much less often. For all speakers, average α was found to be smaller than the typically reported $\alpha \approx 1$. Unfortunately, no values for goodness of fit are given, thus rendering it impossible to evaluate if Zipf's law really holds in all cases.

A fourth and last source of variation discussed here is the medium of the language, in other words, whether the language under consideration is written or spoken. Due to practical limitations, spoken language is studied much less frequently than written language. In addition, written language is often (unconsciously) considered to be simply written down spoken language, which would render any significant differences unlikely. One of the very few studies into Zipf's law in spoken language was performed by Ridley (1982). The transcript that was studied covered an interview with a 25-year old female participant from an unrelated study, in total

2823 words long. He finds Z-β = 1,289, much lower than the expected Z-$\beta \approx 2$. Ridley and Gonzales (1994) looked at smaller samples of only 400 words each. Crucially, they looked at both written and spoken output from the same person. Unfortunately, very few details are given other than that Zipf's law holds.

What this short overview shows is that Zipf's law holds for many different texts from many different sources and languages. It holds for all written language[14] and seems to hold for spoken language as well. But the many different values of α and β that were found throughout the literature cannot directly be compared, because text sizes are different. It therefore remains unclear if any systematic differences exist between texts from different sources. It is also unclear what a reasonable lower limit would be for the number of words that should be included in any analysis. These questions will therefore be addressed in the first part of this chapter.

The second part of the chapter takes a different approach to explore the workings of Zipf's law. The fact that Zipf's law is found in every text in every language means that there are no naturally occurring texts for which Zipf's law does *not* apply. It is therefore an open question if it is at all possible to create a text to which Zipf's law does not apply (while keeping the text readable), and what the result of that would be. To my knowledge, this has not been attempted before. Would readers know what was wrong with the text? How do they evaluate it? This question is addressed in the second part of this chapter.

To summarize, the research questions for the current chapter are as follows:

1 a. How does Zipf's law behave for very small text sizes?
  b. What text sizes are still workable?
  c. Is there a lower limit for which we should state that Zipf's law does not apply?

2. a. Do written and spoken language behave identically, or are there any systematic differences between the two?
  b. Zipf's law is usually studied in large corpora of written language. Can these findings be extended to spoken language? If not, why? What is the difference?

[14] The only exception being Modern Standardized Chinese, discussed in Chapter 2, Section 2.2.1. However, it seems to be that it is not Chinese perse that causes this, but the length of the texts in combination with the limited number of characters. Short texts or phrases probably do follow Zipf's law.

3. a. Is it possible to create a text that does not conform to Zipf's law? What does such a text look like?
   b. How do readers evaluate such a text?

This chapter is divided into two parts. The questions in 1 and 2 are examined in the first part, the questions in 3 are examined in the second part.

3.2 Part I: Small texts, written and spoken

Differences exist in how spontaneous written language is. Some texts – novels, for example – are highly edited and read and re-read until they are considered perfect, while others – such as blog posts – are closer to written down versions of free speech. Similar differences exist within the category of spoken language: words in speeches are often carefully chosen in advance, while words in free conversations are spontaneously chosen on the spot (although some speech planning occurs, of course). Considering these differences, it is possible to construct a continuum from written text to spontaneous speech. In this chapter I use texts from different points on this continuum to examine the difference between written and spoken language. Simultaneously, these texts are used to examine the role of text size.

3.2.1 Methods

The texts that were selected were grouped into the following categories:

Spoken language

1. Face-to-face conversations: unprepared.
2. Spontaneous commentaries on radio or television: unprepared but by trained speakers
3. Radio and television discussions: more or less prepared
4. Sermons/speeches: prepared

Written language:

5. Blogs: informal language
6. News articles: more formal than blogs, but written under time pressure so unlikely to be highly edited
7. Literature: highly edited

An overview of the categories and how they are assumed to be ordered from spontaneous to highly thought through is given in Figure 3.1.

Text selection

Six texts were selected for each category.

All spoken language transcripts were selected from the Dutch (as opposed to Flemish) part of the Corpus of Spoken Dutch (CGN). The face-to-face conversations were selected from CGN component *a*, the spontaneous commentaries from component *f*, discussions from component *i* and sermons and speeches from component *n* and *m*. Per category, the six transcripts from unique speakers with the highest total number of tokens were selected. For the face-to-face conversations, the restriction was added that there should be no more than two speakers per transcript, to prevent small word counts per person. For sermons/speeches, only transcripts with one unique speaker were selected to prevent the inclusion of speeches that take the form of interviews or face-to-face conversations. Details about the transcript are provided in Table 3.1. Per transcript, speech from only one speaker was analysed, which was the speaker who uttered the most tokens.

For all speakers, the language spoken in the transcript, their first language, home language and workplace language were all Standard Dutch. Only for N00063 and

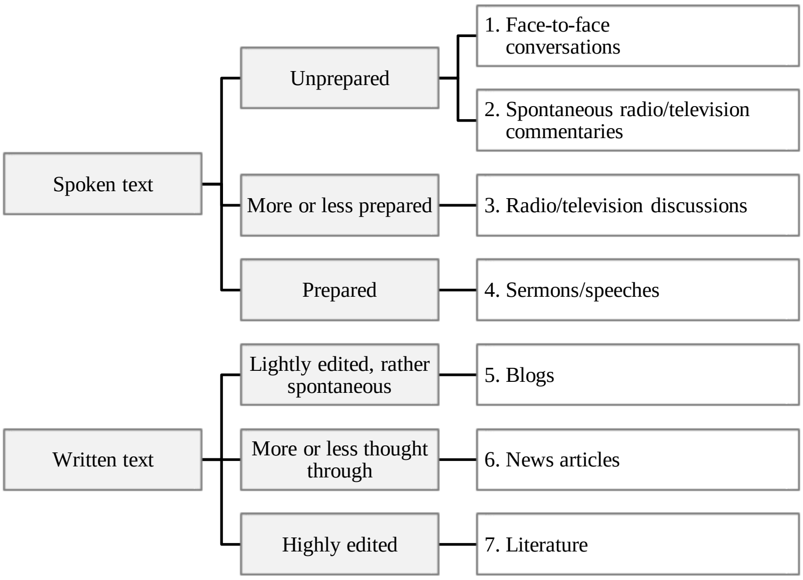

Spoken text
Unprepared
More or less prepared
Prepared
1. Face-to-face conversations
2. Spontaneous radio/television commentaries
3. Radio/television discussions
4. Sermons/speeches
Written text
Lightly edited, rather spontaneous
More or less thought through
Highly edited
5. Blogs
6. News articles
7. Literature

Figure 3.1. Text categories

N00089 first language and home language were unknown. Other speaker details are provided in Table 3.2.

The blogs were selected such that they were all written by different people, on different topics and at least 1000 words. They were found through a number of google searches. The news articles were selected such that they were at least 1000 words long. In two cases, two shorter articles on the same topic from the same source were combined to reach this number, because news articles of more than 1000 words are rare. They were retrieved from the websites or published editions of several major news sources. Literature works were selected that had Dutch as their original language. They were retrieved from DBNL.org, The Gutenberg Project and on one occasion from the website of the author. Text details of the written texts are given in Table 3.3.

For all texts, all punctuation and capitals were removed to prevent the first words of sentences to be treated differently. In the spoken texts, unintelligible speech was transcribed as ‘xxx’, which was removed for further analysis. Unfinished words were removed too. Numerals were removed from the literature texts only, because they were often page numbers. The spoken texts did not contain any numerals. Word counts and type/token counts per text after clean-up are given in Table 3.4 (Results).

Analysis

Type/token-ratio’s (TTR) are calculated as a way to compare word frequencies independent from Zipf’s law. Any differences in TTR between (categories of) texts are expected to be reflected by differences in Zipf's law, since both are calculated based on word frequencies. TTR expresses the number of different words relative to the total number of words. For example, a TTR of 0,33 means that on average, every 30 tokens consist of 10 different types. A TTR of 1 means that every word is unique. A steep slope of Zipf's law, as discussed, occurs when few tokens are used relatively frequently. The same is reflected by a low TTR. A shallow slope, on the other hand, is expected to occur in combination with a high TTR.

Two versions of TTR are calculated, one based on the whole text, and a second based on the first 300 words after the first 50 words. This second TTR was included because TTR based on the whole text is known to be text-size dependent (e.g. Tweedy & Baayen, 1998; Baayen, 2001).

The formula that is used for the Zipf's law-calculations is Zipf-Mandelbrot’s law, Formula 1 in the text box on the second page of this chapter. Both α and Mandelbrot’s β are calculated, here (and elsewhere) dubbed ZM-α and ZM-β. In addition, Zipf's β (Z-β) is calculated, Formula 2 in the aforementioned text box (for more details on the formulas, see Chapter 2, Section 2.2.1). The values of the parameters are calculated through maximum likelihood estimation (Murphy, 2015), after making a first approximation of the parameters through linear regression (Izsák, 2006).

Growth curves are constructed for all three parameters by calculating their values for increasingly large fragments of text, starting at 100 words. For each following analysis, text size is increased with 100 words until full text size no longer allows it or until 5000 words are analysed (only reached by the literary texts). A formula to describe these curves is sought: the type of the distribution is determined based on visual inspection if possible. If no clear distributional type is present, then first, second and third degree polynomial equations are tried for each text (henceforth Model 1, 2 and 3, respectively), and compared using ANOVA's. The corresponding formulas are the following:

3) $y = Ax + B$ (Model 1)
4) $y = Ax^2 + Bx + C$ (Model 2)
5) $y = Ax^3 + Bx^2 + Cx + D$ (Model 3)

The fitted formula will allow for a comparison of the growth curves per category of texts. This way, it can be examined if any systematic differences exist between development of the parameter for the different genres. If a formula can be fitted then it means that all growth curves for that parameter develop in an identical way. If no growth curve can be fitted then this does not automatically mean that the parameter does not stabilize (although that of course can also happen). Rather, it can mean that the fluctuations found for the smallest text samples are too large to fit a uniform formula to. Visual inspection of the growth curves can then still provide valuable insight into the behaviour of the parameter.

Post hoc testing is performed using Tukey's Honest Significant Difference test (Tukey's HSD). This test is similar to a t-test but corrects for multiple comparisons (NIST/SEMATECH *e-Handbook of Statistical Methods*). The aim is to find one model that best describes all growth curves per parameter, after which a comparison between genres can be made.

Based on these growth curves, the minimal workable text size is determined by examining when the shape of the growth curve stabilizes. After determining the minimal workable text size, parameters are compared for fixed numbers of tokens per text.

All analyses are performed in R.

Table 3.1. Details spoken texts

	Rec. ID	CGN Comp.	Content and topic according to the CGN	Word count	Year of rec.	Level of preparation	speakerIDs
1	fn000251	a	Spontaneous conversation; friends	3150	2000	Unprepared	N01001, N01002
	fn000260	a	Spontaneous conversation; parent-child	3156	2000	Unprepared	N01004, N01005
	fn000279	a	Spontaneous conversation; friends	4313	2000	Unprepared	N01010, N01011
	fn000284	a	Spontaneous conversation; friends	3176	2000	Unprepared	N01019, N01020
	fn000300	a	Spontaneous conversation; couple	3374	2000	Unprepared	N01016, N01017
	fn000343	a	Spontaneous conversation; couple	2888	2000	Unprepared	N01028, N01029
2	fn007440	i	Radio: Langs de lijn; sports; EK; soccer; Netherlands-Denmark	4529	2000	Unprepared	N03104, N03113
	fn007484	i	Radio: Langs de lijn; sports; EK; swimming	1040	1999	Unprepared	N03096
	fn008691	i	Television: Na Elven; sports; basketball	1033	2002	Unprepared	N04067
	fn008781	i	Television: Studio Sport; triathlon; Olympics	1406	2000	Unprepared	N04086
	fn008958	i	Television: Studio Sport; luge; Winter Olympics	1064	2002	Unprepared	N04087
	fn007056	i	Radio: Bijlmer Enquête; Bijlmer disaster; parliamentary inquiry	1532	1999	Unprepared	N04036
3	fn007147	f	Radio: Het debat; Christianity; norms and values; capitalism	2986	1998	Somewhat prepared	N00168, N00337, N00339, N03081
	fn007361	f	Radio: Punch; politics; PvdA	3995	2000	Somewhat prepared	N03144, N03362, N03363
	fn007367	f	Radio: Zuiderlicht (VPRO aan de Amstel); science; genetic research; recombination processes	3506	1999	Somewhat prepared	N03273, N03371
	fn007494	f	Radio: Vroege vogels; nature; bats	3196	2000	Somewhat prepared	N03046, N03446, N03447, N03467
	fn007504	f	Radio: De tafel van Pam; politics; the extreme right	5408	1998	Somewhat prepared	N03308, N03309, N03310, N03311, N03479, N03480
	fn007568	f	Radio: Nieuwsshow; soccer; soccer stadions	3020	1998	Somewhat prepared	N03172, N03571, N03572
4	fn000271	m	Eucharist: wedding service	1502	2000	Prepared	N00390
	fn000057	n	Presentation	4140	1999	Prepared	N00055
	fn000064	n	Presentation	4645	1999	Prepared	N00051
	fn000065	n	Presentation	5132	1999	Prepared	N00050
	fn000068	n	Presentation	3394	1999	Prepared	N00063
	fn000080	n	presentation	3679	1999	Prepared	N00089

Table 3.2. Speaker details

	Rec. ID	Speaker ID	Age	Sex	Birth region	Residence region	Level of education	Occupation
1	fn000251	N01001	45-55	f	Flanders: Antwerpen and Vlaams-Brabant	Gelders rivierengebied	High	Professor
	fn000260	N01004	25-34	f	Noord-Holland excl. West-Friesland	Gelders rivierengebied	High	Doctoral student
	fn000279	N01010	25-34	m	Noord-Brabant	Gelders rivierengebied	High	Doctoral student
	fn000284	N01019	35-44	m	Noord-Brabant	West-Utrecht, incl. the city of Utrecht	High	Consultant
	fn000300	N01017	25-34	m	Noord-Holland excl. West-Friesland	West-Utrecht, incl. the city of Utrecht	High	Engineer
	fn000343	N01029	Over 55	m	Gelders rivierengebied	Gelders rivierengebied	High	Planner
2	fn007440	N04036	45-55	m	Zuid-Holland excl. Goeree Overflakee	Unknown	Unknown	Commentator
	fn007484	N03113	Unkn.	m	Noord-Holland excl. West-Friesland	Noord-Holland excl. West-Friesland	High	Reporter
	fn008691	N03096	Unkn.	m	Unknown	Unknown	Unknown	Reporter
	fn008781	N04067	25-34	m	Unknown	Unknown	Unknown	Commentator
	fn008958	N04086	25-34	m	Drenthe	Noord-Holland excl. West-Friesland	Unknown	Commentator
	fn007056	N04087	45-55	m	Unknown	Veluwe up to the river IJssel	High	Commentator
3	fn007147	N00337	Over 55	f	Noord-Holland excl. West-Friesland	Oost-Utrecht excl. the city of Utrecht	High	Minister
	fn007361	N03362	25-34	m	Noord-Brabant	Unknown	High	Entrepreneur
	fn007367	N03371	35-44	m	Unknown	Unknown	High	Professor
	fn007494	N03467	Unkn.	m	Unknown	Unknown	High	Biologist
	fn007504	N03480	35-44	m	Unknown	Unknown	Unknown	Publicist
	fn007568	N03571	Unkn.	m	Unknown	Noord-Holland excl. West-Friesland	Unknown	Unspecified
4	fn000271	N00390	Over 55	m	Limburg	Gelders rivierengebied	High	Clergyman
	fn000057	N00055	Over 55	m	Outside of The Netherlands and Flanders	West-Utrecht, incl. the city of Utrecht	High	Physicist
	fn000064	N00051	45-55	m	Gelders rivierengebied	West-Utrecht, incl. the city of Utrecht	Middle	Consultant
	fn000065	N00050	45-55	m	Gelders rivierengebied	Oost-Utrecht excl. the city of Utrecht	High	Professor
	fn000068	N00063	Unkn.	f	Unknown	NL, unknown	Unknown	Unspecified
	fn000080	N00089	Unkn.	m	Unknown	NL, unknown	Unknown	unspecified

Table 3.3. Details written texts. All digital sources were accessed on 9-8-2016.

	Author	Title	Pub. date	Source
5	Elja Daae	Ondernemerslessen	9-8-2016	www.eljadaae.nl
	Dina	Twee maanden mama	20-6-2016	www.dinastie.nl
	Hester	In vogelvlucht en snapshots	29-7-2016	www.hesterly.nl
	Sarra	Zakgeld, hoe ga je hiermee om?	17-2-2016	lepetitfavorite.com
	Thamar Kempees	Tattootijd: Art Collective Almelo	27-7-2016	www.thamarkempees.nl
	Aïsha	Toen ik ging trouwen, maar toch niet	29-3-2016	www.leesjeblij.nl
6	Eva Cukier; Marc Leijendekker	Ergernis verenigt Poetin en Erdogan; ‘Ban Erdogan uit de Duitse klaslokalen’	9-8-2016	*NRC*, p. 1, 13
	Joris Belgers	Bovenop het nieuws, zonder journalisten	26-7-2016	www.trouw.nl
	Tamar Stelling	Echt vrije seks vind je ver beneden peil	7-7-2016	decorrespondent.nl
	John Volkers; "Redactie"	Einde Spelen voor Yuri van Gelder na alcoholgebruik; In beeld: Hoe Yuri van Gelder opkwam, viel, opkrabbelde en weer viel	9-8-2016	www.volkskrant.nl
	Roelf Jan Duin & Hanneke Keultjes	Rio ziet Eurlings op zijn charmantst	6-8-2016	*Het Parool*, p. 6
	Frans Verhagen	Het land van Nixon	3-8-2016	*De Groene Amsterdammer*, p. 9-10
7	F. Bordewijk	Karakter	1938	www.dbnl.org
	Hildebrand (Nicolaas Beets)	Camera Obscura	1917	The Project Gutenberg
	Multatuli (Eduard Douwes Dekker)	Max Havelaar	1860	The Project Gutenberg
	Stefan Nieuwenhuis	Ik ben omringd door debielen en ik voel me goed	2012	stefannieuwenhuis.nl
	Frederik van Eeden	De kleine Johannes	1887	The Project Gutenberg
	Joost van den Vondel	Lucifer	1654	www.dbnl.org

3.2.2 Results

Type/token ratio

Number of types, tokens and both size dependent and size independent type/token-ratio (TTR) per individual text can be found in Table 3.4. Mean number of types, tokens and TTR per category can be found in Table 3.5.

A one-way analysis of variance shows that the differences between the size-dependent TTR’s is significant (F(6,35) = 36,863, *p* < 0,001). A Tukey’s Honestly Significant Difference (Tukey’s HSD) post-hoc test reveals that conversations do

Table 3.4. Types, tokens and type/token ratio per text. Size-dependent TTR was calculated using full text size. Size-independent TTR was calculated using the number of types in the first 300 tokens after the first 50 tokens.

	Text	**Total no. Types**	**Total no. Tokens**	**TTR (size-dependent)**	**TTR (size-independent)**
1	fn000251	419	1716	0,244	0,450
	fn000260	483	2215	0,218	0,440
	fn000279	627	3221	0,195	0,453
	fn000284	502	2017	0,249	0,437
	fn000300	419	1646	0,255	0,457
	fn000343	397	1449	0,274	0,460
2	fn007056	477	1520	0,314	0,507
	fn007440	537	2120	0,253	0,523
	fn007484	351	1033	0,340	0,493
	fn008691	363	1030	0,352	0,573
	fn008781	437	1405	0,311	0,497
	fn008958	317	1063	0,298	0,450
3	fn007147	363	1036	0,350	0,497
	fn007361	550	1873	0,294	0,497
	fn007367	607	2254	0,269	0,527
	fn007494	569	2004	0,284	0,473
	fn007504	656	2702	0,243	0,510
	fn007568	441	1803	0,245	0,343
4	fn000057	752	4084	0,184	0,453
	fn000064	1089	4476	0,243	0,540
	fn000065	941	5037	0,187	0,483
	fn000068	722	3363	0,215	0,473
	fn000080	906	3640	0,249	0,480
	fn000271	456	1476	0,309	0,540
5	Aïsha	456	1124	0,406	0,620
	Dina	380	1019	0,373	0,533
	Elja Daae	406	1008	0,403	0,580
	Hester	581	1319	0,440	0,597
	Sarra	405	1210	0,335	0,483
	Thamar Kempees	469	1160	0,404	0,600
6	Cukier; Leijendekker	645	1446	0,446	0,620
	Verhagen	1131	2826	0,400	0,650
	Volkers; "Redactie"	540	1531	0,353	0,573
	Belgers	617	1646	0,375	0,553
	Duin & Keultjes	570	1243	0,459	0,620
	Stelling	877	2038	0,430	0,633
7	Bordewijk	10237	87777	0,117	0,560
	Hildebrand	18526	157739	0,117	0,643
	Multatuli	13799	126979	0,109	0,600
	Nieuwenhuis	6789	52474	0,129	0,627
	Van den Vondel	4426	20347	0,218	0,643
	Van Eeden	5240	37848	0,138	0,563

not differ from discussions or speeches, commentaries do not differ from discussions or speeches and blogs do not differ from articles. All other differences in TTR are significant. The highest TTR (0,411) is found for articles, the lowest TTR (0,138) is found for literature. For *p*-values, see Table 3.6.

The differences between the size-independent TTR's are also significant, as shown by a one-way analysis of variance (F(6,35) = 13,34; $p < 0{,}001$). A Tukey's HSD post-hoc test reveals that there are clear differences between written and spoken texts. No differences exist between conversations, commentaries, discussions and speeches, or between blogs, articles and literature. Significant differences do exist between blogs and conversations, discussions and speeches, between articles and all four spoken categories and between literature and all four spoken categories. Generally speaking, size-independent TTR's are smaller for spoken texts than for written texts, reflecting a smaller vocabulary in spoken language than in written language. For *p*-values, see Table 3.7.

Growth curves

Zipf-Mandelbrot's α

The growth curves of the ZM-α values for the first 2000 tokens are shown in Figure 3.2. Values for ZM-α range from ZM-α = 0,3 to ZM-α = 1,3. As discussed above, a higher ZM-α reflects a less diverse vocabulary. The longer the text, the bigger the chance that any new word has been encountered before, thus increasing the value of ZM-α. In all cases, ZM-α increases when the analysed number of tokens increases. The growth curves seem to follow a logarithmic curve. This was confirmed by plotting the growth curves on logarithmic x-axes (Figure 3.3): a typical property of a logarithmic distribution is that it becomes a linear distribution on logarithmic scales (or after logarithmic transformation of the values).

The formula that corresponds to the fitted logarithmic model is

6) $\alpha(x) = A \log x + B$

where x is the number of tokens and A and B are the parameters of the model. The mean parameters values of the fitted model and the corresponding mean adjusted R^2-values are given in Table 3.8. Values per text are given in Table 3.9.

Table 3.5. Mean types, tokens and type/token ratio per text. Size-dependent TTR was calculated using full text size. Size-independent TTR was calculated using the number of types in the first 300 tokens after the first 50 tokens.

Category	**No. types**	**No. tokens**	**TTR (size dependent)**	**No. types in tokens 51-350**	**TTR (size-independent)**
1. conversations	475	2044	0,239	135	0,450
2. commentaries	414	1362	0,311	152	0,507
3. discussions	531	1945	0,281	142	0,475
4. speeches	811	3679	0,231	149	0,495
5. blogs	450	1140	0,394	170	0,569
6. articles	730	1788	0,411	183	0,608
7. literature	9836	80527	0,138	182	0,606

Table 3.6. TTR (size-dependent) p-values (Tukey's HSD post-hoc test)

	7. literature	**6. articles**	**5. blogs**	**4. speeches**	**3. discussions**	**2. commentaries**
1. conversations	*0,001*	*0,000*	*0,000*	1,000	0,512	*0,038*
2. commentaries	*0,000*	*0,001*	*0,012*	*0,015*	0,814	
3. discussions	*0,000*	*0,000*	*0,000*	0,306		
4. speeches	*0,003*	*0,000*	*0,000*			
5. blogs	*0,000*	0,987				
6. articles	*0,000*					

Table 3.7. TTR (size-independent) p-values (Tukey's HSD post-hoc test)

	7. literature	**6. articles**	**5. blogs**	**4. speeches**	**3. discussions**	**2. commentaries**
1. conversations	*0,000*	*0,000*	*0,005*	0,541	0,949	0,263
2. commentaries	*0,006*	*0,004*	0,197	0,999	0,841	
3. discussions	*0,000*	*0,000*	*0,009*	0,981		
4. speeches	*0,001*	*0,001*	*0,070*			
5. blogs	0,745	0,694				
6. articles	1,000					

Table 3.8. Mean curve fitting outcomes per category for ZM-α

Text	**A**	**B**	**R^2**
1. Conversations	0,211	-0,393	0,957
2. Commentaries	0,182	-0,340	0,971
3. Discussions	0,178	-0,217	0,954
4. Sermons/speeches	0,139	-0,029	0,942
5. Blogs	0,190	-0,435	0,943
6. Articles	0,137	-0,183	0,970
7. Literature	0,126	-0,082	0,948

Table 3.9. Curve fitting outcomes per text for ZM-*α*

	Text	**A**	**B**	**R^2**
1	fn000251	0,224	-0,443	0,942
	fn000260	0,204	-0,253	0,960
	fn000279	0,185	-0,202	0,944
	fn000284	0,214	-0,544	0,975
	fn000300	0,231	-0,562	0,966
	fn000343	0,211	-0,356	0,955
2	fn007056	0,146	-0,172	0,990
	fn007440	0,213	-0,581	0,985
	fn007484	0,176	-0,200	0,899
	fn008691	0,198	-0,491	0,994
	fn008781	0,133	-0,086	0,985
	fn008958	0,227	-0,508	0,978
3	fn007147	0,204	-0,380	0,989
	fn007361	0,203	-0,402	0,986
	fn007367	0,117	0,147	0,912
	fn007494	0,197	-0,360	0,983
	fn007504	0,177	-0,257	0,960
	fn007568	0,171	-0,051	0,891
4	fn000057	0,150	-0,167	0,970
	fn000064	0,129	-0,004	0,974
	fn000065	0,165	-0,079	0,873
	fn000068	0,156	-0,194	0,986
	fn000080	0,075	0,495	0,927
	fn000271	0,155	-0,226	0,924
5	Aïsha	0,199	-0,516	0,957
	Dina	0,202	-0,494	0,973
	Elja Daae	0,169	-0,327	0,989
	Hester	0,176	-0,383	0,884
	Sarra	0,179	-0,241	0,901
	Thamar Kempees	0,217	-0,648	0,955
6	Belgers	0,150	-0,250	0,985
	Cukier; Leijendekker	0,117	-0,084	0,968
	Duin Keultjes	0,175	-0,414	0,962
	Stelling	0,119	-0,112	0,969
	Verhagen	0,097	0,053	0,972
	Volkers; Redactie	0,162	-0,292	0,963
7	Bordewijk	0,104	0,105	0,928
	Hildebrand	0,130	-0,148	0,947
	Multatuli	0,139	-0,145	0,986
	Nieuwenhuis	0,146	-0,194	0,947
	Van den Vondel	0,132	-0,206	0,990
	Van Eeden	0,104	0,096	0,892

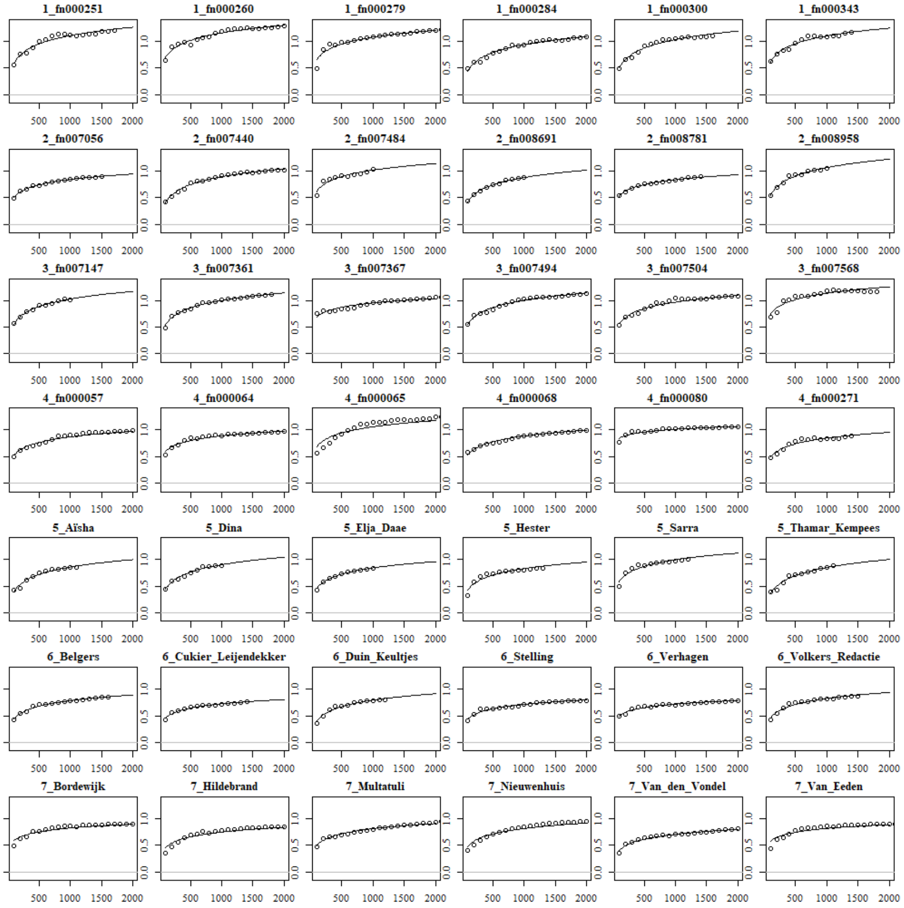


Figure 3.2. ZM-*α* growth curves (first 2000 tokens) with fitted curve. On x-axis: text size; on y-axis: ZM-*α* value.

The average adjusted R^2 for all texts was 0,955 (SD = 0,034), ranging from 0,873 (speech fn000065) to 0,994 (commentary fn008691). This can be considered a good fit.

Significant differences exist between categories for parameter A (Formula 6), as a one-way ANOVA shows (F(6, 35) = 8,172, *p* < 0,001). The parameter A determines the rate at which the value of ZM-*α* grows. A Tukey's HSD post-hoc test reveals that conversations have a higher A than speeches (*p* < 0,001), commentaries have a higher A than literature (*p* = 0,020), discussions have a higher A than literature (*p* = 0,036), and blogs have a higher A than speeches, articles and literature (*p* = 0,038, *p* = 0,028 and *p* = 0,005, respectively; see Table 3.10 for all *p*-values). It

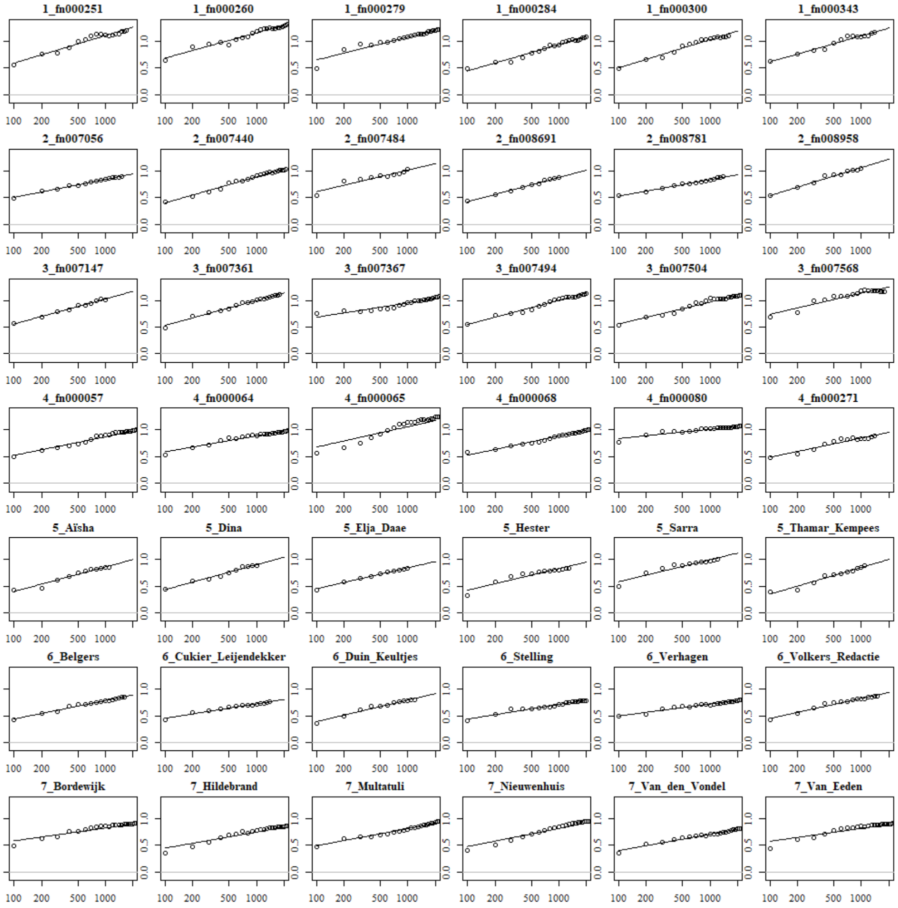


Figure 3.3. ZM-α growth curves (first 2000 tokens) with fitted curve on logarithmic x-axes. On x-axis: text size; on y-axis: ZM-α value.

thus seems to be the case that a higher A is associated with more spontaneous language output. It is striking that blogs, in this respect, are closer to conversations and commentaries than to other written speech output. These outcomes thus mean that ZM-α grows faster for most categories of spoken language than for most categories of written language, except for blogs. The growth rate of ZM-α reflects the chance that, while analysing increasingly large chunks of text, a newly encountered word has been encountered before: this chance is higher in spoken language and blogs than in written language.

Significant differences exist for parameter B too (F(6,35) = 3,927, *p* = 0,004). Parameter B determines the intercept of the curve, in other words, the value of ZM-α

for very small samples. A Tukey's HSD post-hoc test reveals that this difference is due to speeches having a higher intercept than conversations (p = 0,034) and blogs (p = 0,013); and literature having a higher intercept than blogs (p = 0,043). See Table 3.11 for all p-values.

Table 3.10. Parameter A, ZM-α growth curves, p-values (Tukey's HSD post-hoc test)

	7. literature	**6. articles**	**5. blogs**	**4. speeches**	**3. discussions**	**2. commentaries**
1. conversations	*<0,001*	*0,001*	0,842	*0,001*	0,389	0,537
2. commentaries	*0,020*	0,095	0,998	0,123	1	
3. discussions	*0,036*	0,157	0,987	0,200		
4. speeches	0,985	1	*0,038*			
5. blogs	*0,005*	*0,028*				
6. articles	0,994					

Table 3.11. Parameter B, ZM-α growth curves, p-values (Tukey's HSD post-hoc test)

	7. literature	**6. articles**	**5. blogs**	**4. speeches**	**3. discussions**	**2. commentaries**
1. conversations	0,102	0,496	1	*0,034*	0,689	1
2. commentaries	0,262	0,792	0,976	0,103	0,922	
3. discussions	0,882	1	0,452	0,622		
4. speeches	0,999	0,803	*0,013*			
5. blogs	*0,043*	0,285				
6. articles	0,968					

Zipf-Mandelbrot's β

The growth curves for ZM-β for the first 2000 tokens are displayed in Figure 3.4. The picture is clearly messier than it was for ZM-α. Values cover a wide range, going from ZM-$\alpha = -0{,}8$ to ZM-α = 10,5. For some texts ZM-β remains relatively stable as text size increases, for others it increases, and in some cases, it decreases. Some curves remain relatively linear, others seem to display two-phase behaviour: ZM-β grows linearly until it levels off quite abruptly.

Visual inspection does not give one uniform distribution that seems to fit all growth curves. Therefore, an attempt to curve fitting was performed following the method described above: first, second and third degree polynomial equations are tried for each text. The result of this per text can be found in Table 3.12. Curve fitting following this method did not work for all curves. Large differences between texts were found: for Model 1, the smallest adjusted R^2 for any text was as low as −0,079 (blog by Hester), the highest was 0,967 (conversation fn000284). For Model 2, the spread reached from -0,067 (blog by Hester) to 0,974 (conversation fn000284); and

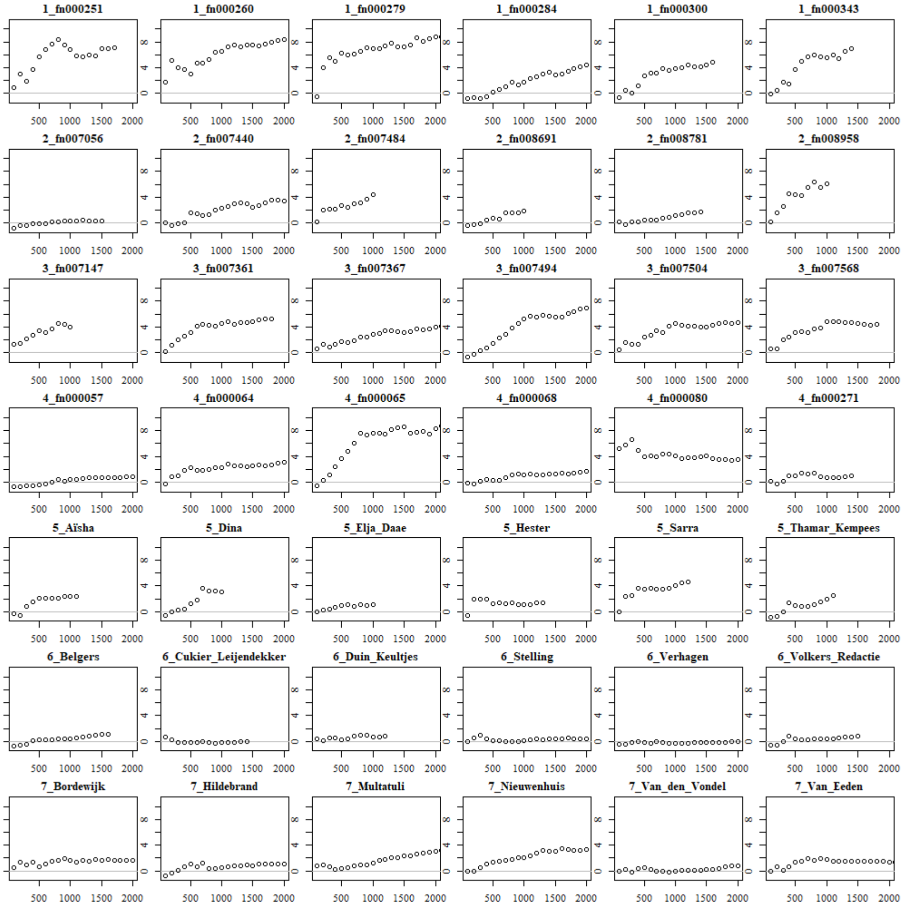


Figure 3.4. Growth curves for ZM-β (first 2000 tokens). On x-axis: text size; on y-axis: ZM-β values.

for Model 3 from 0,136 (article by Stelling) to 0,981 (discussion fn007361). In other words, not a single model worked to a reasonable extent (in terms of consistent and reasonably high R^2-values) for all texts. Neither could a pattern be discovered such that some models fitted better to some categories than others (probably also as a result of the apparent two-phase behaviour of some curves). It was therefore decided to not statistically compare groups based on the parameters of the polynomial functions.

Table 3.12. Curve fitting outcomes for Zipf-Mandelbrot's β

	Text	R^2 poly1	R^2 poly2	R^2 poly3	p 1 vs. 2	p 2 vs. 3
1	fn000251	0,395	0,678	0,805	0,002	0,007
	fn000260	0,856	0,881	0,875	0,035	0,873
	fn000279	0,757	0,835	0,849	0,001	0,065
	fn000284	0,967	0,974	0,972	0,032	0,932
	fn000300	0,815	0,942	0,948	0,000	0,130
	fn000343	0,810	0,917	0,914	0,002	0,453
2	fn007056	0,775	0,950	0,946	0,000	0,988
	fn007440	0,869	0,918	0,913	0,002	0,966
	fn007484	0,827	0,806	0,931	0,730	0,010
	fn008691	0,948	0,944	0,946	0,495	0,311
	fn008781	0,921	0,931	0,948	0,120	0,060
	fn008958	0,829	0,933	0,925	0,008	0,619
3	fn007147	0,890	0,937	0,946	0,033	0,201
	fn007361	0,758	0,923	0,981	0,000	0,000
	fn007367	0,922	0,959	0,957	0,000	0,635
	fn007494	0,889	0,969	0,968	0,000	0,648
	fn007504	0,656	0,909	0,919	0,000	0,057
	fn007568	0,712	0,958	0,955	0,000	0,701
4	fn000057	0,762	0,958	0,974	0,000	0,000
	fn000064	0,853	0,904	0,920	0,000	0,005
	fn000065	0,355	0,666	0,914	0,000	0,000
	fn000068	0,949	0,959	0,961	0,006	0,114
	fn000080	0,262	0,685	0,735	0,000	0,011
	fn000271	0,155	0,515	0,605	0,009	0,091
5	Aïsha	0,720	0,898	0,890	0,004	0,550
	Dina	0,859	0,861	0,932	0,326	0,027
	Elja Daae	0,738	0,925	0,915	0,003	0,685
	Hester	-0,079	-0,067	0,390	0,315	0,017
	Sarra	0,646	0,737	0,918	0,064	0,002
	Thamar Kempees	0,792	0,775	0,852	0,572	0,057
6	Belgers	0,909	0,912	0,949	0,249	0,007
	Cukier_Leijendekker	0,215	0,703	0,815	0,001	0,020
	Duin_Keultjes	0,477	0,447	0,446	0,518	0,350
	Stelling	-0,015	0,121	0,136	0,069	0,271
	Verhagen	0,689	0,726	0,731	0,044	0,251
	Volkers_Redactie	0,480	0,532	0,698	0,143	0,019
7	Bordewijk	0,300	0,294	0,749	0,466	0,000
	Hildebrand	0,717	0,811	0,830	0,000	0,015
	Multatuli	0,768	0,927	0,927	0,000	0,260
	Nieuwenhuis	0,465	0,859	0,981	0,000	0,000
	Van_den_Vondel	0,923	0,926	0,938	0,101	0,003
	Van_Eeden	0,303	0,499	0,587	0,000	0,002

Zipf's β

The procedure followed for ZM-α and for ZM-β was once more repeated to examine the growth curves for Z-β. The growth curves for Z-β for the first 2000 tokens are displayed in Figure 3.5. Z-β either decreases slightly when text size grows, or it stays relatively stable. The result of the curve fitting per text can be found in Table 3.13.

The average adjusted R^2 for all texts was 0,532 (SD: 0,274) for Model 1, 0,658 (SD: 0,282) for Model 2, and 0,730 (SD: 0,235) for Model 3. But as with ZM-β, large differences between texts were found: for Model 1, the smallest adjusted R^2 for any text was as low as −0,020 (speech fn000080), the highest was 0,876 (commentary fn008691). For Model 2, the spread ranged from −0,012 (commentary fn007484) to 0,978 (commentary fn008691); and for Model 3 from −0,157 (commentary fn007484) to 0,978 (commentary fn008691). In other words, not a single model worked to a reasonable extent for all texts (by which a consistently high R^2 for all texts is meant) and no pattern was discovered. Values were therefore not statistically compared.

Looking at the growth curves it seems to be the case that for most texts, the first two data points are not in line with the others. It was therefore also attempted to perform curve fitting while excluding these two points. The results were not much better: R^2 values still ranged from -0,093 to 0,913 for Model 1, from -0,119 to 0,970 for Model 2 and from -0,099 to 0,994 for Model 3. Mean values were 0,546 (SD: 0,274) for

Model 1, 0,630 for Model 2 (SD: 282) and 0,714 (SD: 0,235) for Model 3. So still, no single model worked for all texts.

It seems likely that fit of the models will increase if larger texts are examined. Unfortunately, larger texts were not available from the current sources. All curves appear to flatten out after some initial variation – a pattern very different from the growth curves for Mandelbrot's β. Unfortunately, small text sizes prevent statistically testing this hypothesis.

Table 3.13. Curve fitting outcomes for Zipf's *β*

	Text	**R^2 poly1**	**R^2 poly2**	**R^2 poly3**	**p 1 vs. 2**	**p 2 vs. 3**
1	fn000251	0,373	0,627	0,760	0,005	0,011
	fn000260	0,391	0,371	0,432	0,553	0,098
	fn000279	0,673	0,773	0,772	0,001	0,363
	fn000284	0,644	0,667	0,647	0,151	0,960
	fn000300	0,786	0,920	0,916	0,000	0,567
	fn000343	0,702	0,685	0,654	0,558	0,971
2	fn007056	0,699	0,790	0,829	0,024	0,081
	fn007440	0,642	0,847	0,924	0,000	0,000
	fn007484	0,087	-0,012	-0,157	0,659	0,739
	fn008691	0,876	0,978	0,978	0,000	0,352
	fn008781	0,861	0,918	0,920	0,011	0,285
	fn008958	0,510	0,801	0,933	0,009	0,009
3	fn007147	0,696	0,895	0,880	0,005	0,709
	fn007361	0,434	0,627	0,775	0,008	0,005
	fn007367	0,601	0,747	0,744	0,002	0,397
	fn007494	0,308	0,541	0,554	0,005	0,237
	fn007504	0,454	0,642	0,787	0,001	0,000
	fn007568	0,088	0,104	0,365	0,276	0,018
4	fn000057	0,572	0,809	0,869	0,000	0,000
	fn000064	0,772	0,911	0,910	0,000	0,579
	fn000065	0,253	0,391	0,613	0,001	0,000
	fn000068	0,344	0,468	0,666	0,007	0,000
	fn000080	-0,020	0,306	0,334	0,000	0,135
	fn000271	0,484	0,559	0,857	0,110	0,001
5	Aïsha	0,871	0,974	0,974	0,000	0,389
	Dina	0,747	0,938	0,954	0,001	0,114
	Elja Daae	0,575	0,691	0,870	0,086	0,017
	Hester	0,276	0,517	0,640	0,029	0,064
	Sarra	0,251	0,536	0,544	0,026	0,311
	Thamar Kempees	0,587	0,767	0,784	0,023	0,242
6	Belgers	0,258	0,547	0,891	0,008	0,000
	Cukier_Leijendekker	0,316	0,311	0,310	0,359	0,344
	Duin_Keultjes	0,797	0,884	0,891	0,017	0,235
	Stelling	0,495	0,508	0,484	0,239	0,665
	Verhagen	0,554	0,631	0,639	0,018	0,225
	Volkers_Redactie	0,523	0,505	0,791	0,488	0,002
7	Bordewijk	0,440	0,571	0,820	0,000	0,000
	Hildebrand	0,554	0,586	0,640	0,035	0,007
	Multatuli	0,791	0,932	0,931	0,000	0,455
	Nieuwenhuis	0,615	0,873	0,947	0,000	0,000
	Van_den_Vondel	0,854	0,899	0,899	0,000	0,297
	Van_Eeden	0,621	0,616	0,680	0,603	0,002

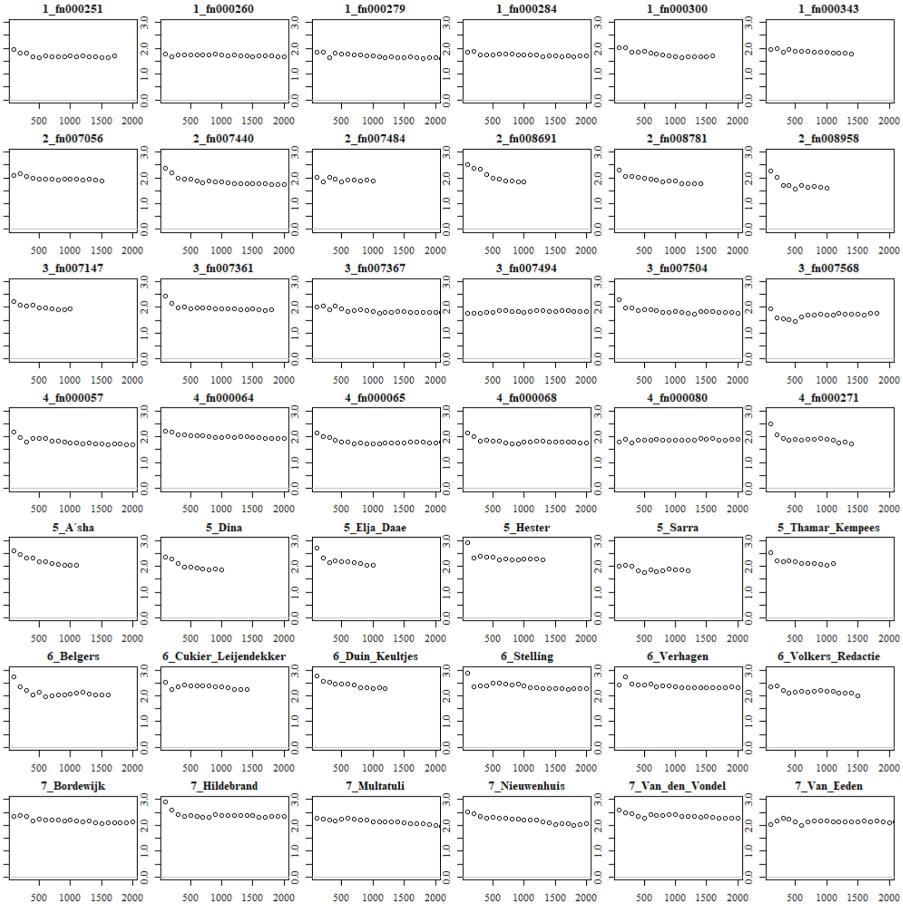


Figure 3.5. Growth curves for Z-β (first 2000 tokens). On x-axis: text size; on y-axis: Z-β values.

Minimal workable text size

Above, it was found that for the text sizes examined here, the value of ZM-α grows as text size increases. At first it increases rapidly, until it continues at a seemingly constant rate (following a logarithmic distribution). This means that the difference in ZM-α between any point 2 and 3 minus the difference in ZM-α between any point 1 and 2 approaches 0. In short:

7) $$\Delta\Delta_{\text{ZM-}\alpha} = \Delta_{\text{ZM-}\alpha^{x+2} - \text{ZM-}\alpha^{x+1}} - \Delta_{\text{ZM-}\alpha^{x+1} - \text{ZM-}\alpha^{x}} \rightarrow 0$$

The point at which this happens can be considered the smallest workable text size.

From this point onwards, changes in ZM-α can be considered to be due to the combination of text size and the specific characteristics of the text as a whole, and not due to idiosyncrasies such as the introduction of a new topic.

The development of $\Delta\Delta_{\text{ZM-}\alpha}$ per text can be seen in Figure 3.6. It should be kept in mind that some texts that were studied here are rather small, so not all texts are large enough to draw definite conclusions. Tentatively, however, it can be seen that at least 1500 tokens are needed for $\Delta\Delta_{\text{ZM-}\alpha}$ to approach 0. Some texts approach 0 earlier (e.g. the article by Cukier and Leijendekker), some need slightly more words (e.g. speech fn000065). It seems safe to conclude that analyses of ZM-α can safely be performed for texts of 2000 words or more.

Similarly to $\Delta\Delta_{\text{ZM-}\alpha}$, $\Delta\Delta_{\text{ Z-}\beta}$ can be calculated[15]:

8) $\Delta\Delta_{Z\text{-}\beta} = \Delta_{Z\text{-}\beta^{x+2} - Z\text{-}\beta^{x+1}} - \Delta_{Z\text{-}\beta^{x+1} - Z\text{-}\beta^{x}} \rightarrow 0$

The development of $\Delta\Delta_{\text{Z-}\beta}$ per text can be seen in Figure 3.7. For those texts that are long enough, it seems to be the case that the same conclusions can be drawn as for ZM-α. At least 1500 tokens are necessary for $\Delta\Delta_{\text{Z-}\beta}$ to approach 0. To be on the safe side, it seems like a wise decision to take slightly larger text samples. 2000 tokens can be considered a safe sample size for analyses of Z-β.

It is important to note that this does not mean that ZM-α or Z-β does not change for texts of larger sizes. What it does mean is that the rate at which ZM-α or Z-β changes stabilizes. It is therefore important to always compare texts of identical text size. This identical text size should thus be at least 1500 words or more.

Parameters for fixed text sizes

Now that the minimal workable text size for Zipf's law is known, we can compare the values of the parameters for fixed text sizes to see if any differences exist between the different categories of texts. Unfortunately, not all text samples reach the required 2000 words. Therefore, it was decided to perform the analysis for 1000 words (reached by all texts), 1500 words and 2000 words to see if any pattern arises. The results from the 1000-word analyses should thus be interpreted with caution, since values can fluctuate if texts had been larger.

[15] These analyses were not performed for ZM-β because of the high diversity in the growth curves and absence of any asymptotes.

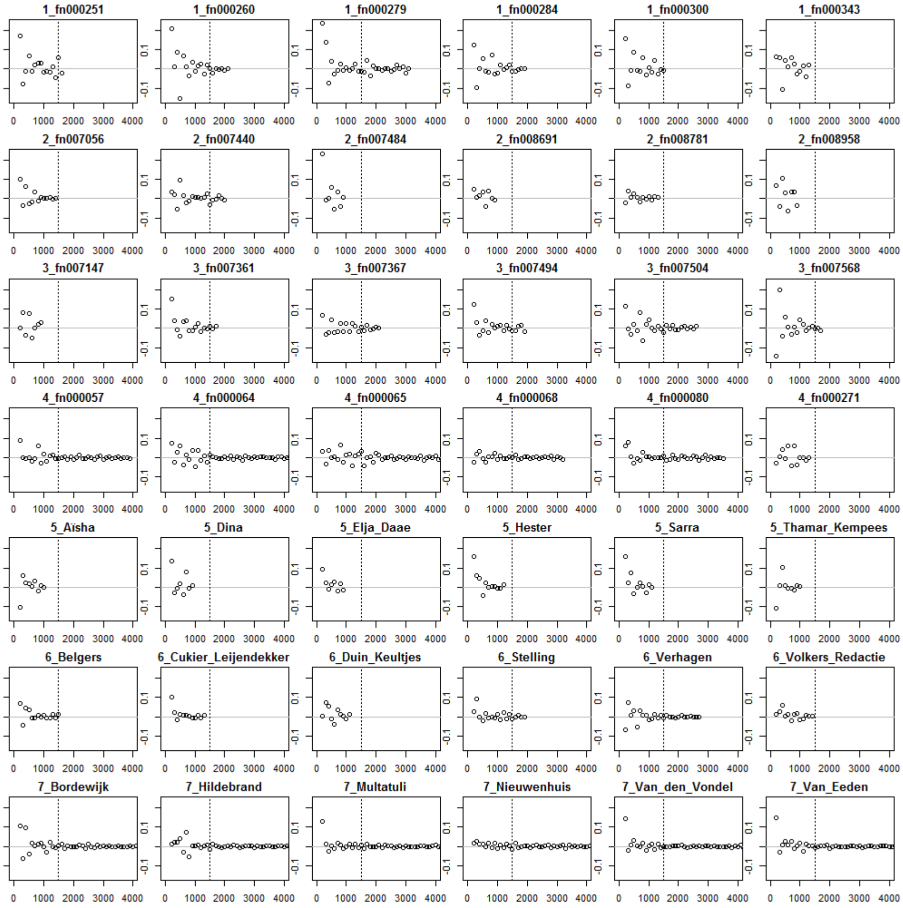


Figure 3.6. ΔΔZM-*α* per text. The vertical line indicates x = 1500. On x-axis: text size. On y-axis: ΔΔZM-*α* values.

Values per text for ZM-*α* and ZM-*β* and for Z-*β* per text are given in Table 3.14. Mean values per category are given in Table 3.15.

A one-way analysis of variance shows that the differences between the ZM-*α*-values per category are significant for all token counts[16] (1000 tokens: F(6,35) = 13,389, *p* < 0,001; 1500 tokens: F(5,21) = 12,727, *p* < 0,001; 2000 tokens: F(5,14) = 10,277,

[16] For completeness, statistical tests for all token counts are reported. However, it should be kept in mind that the number of texts that are involved were small to start with, and even

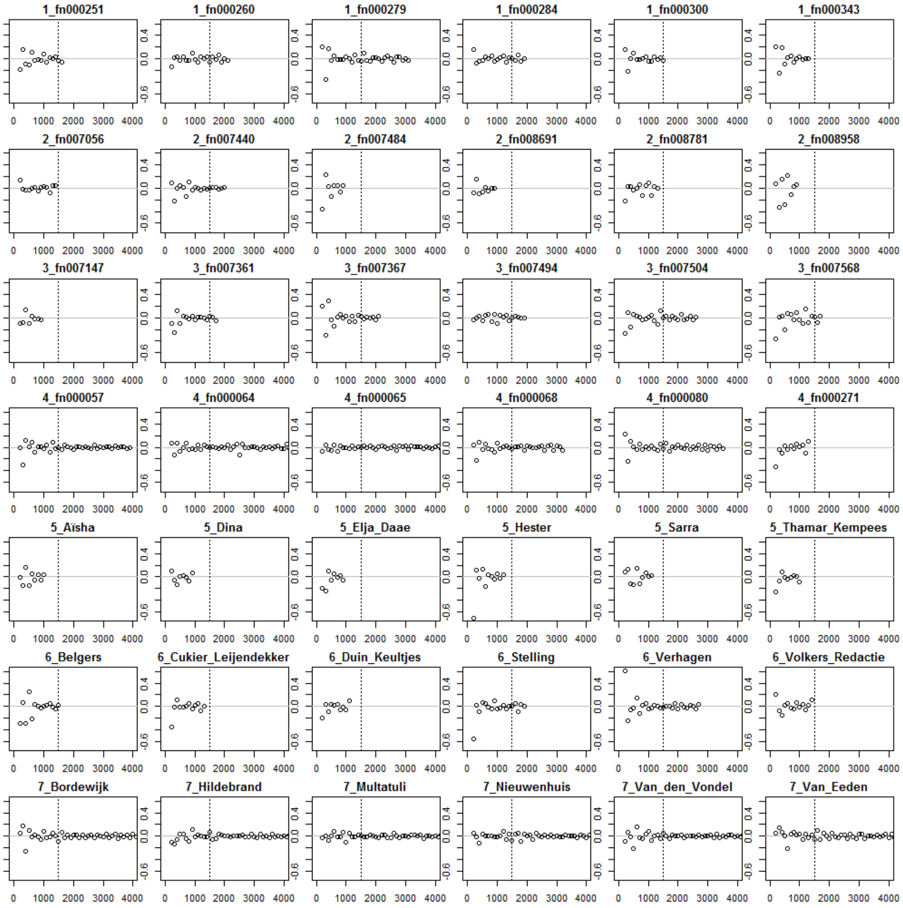


Figure 3.7. ΔΔZ-β per text. The vertical line indicates x = 1500. On x-axis: text size; on y-axis: ΔΔZ-β values.

p < 0,001). A Tukey's Honestly Significant Difference (Tukey's HSD) post-hoc test reveals that no differences exist within the group of written texts for any token count. Within the group of spoken texts, conversations display significantly higher ZM-α-values than discussions for 1000 and 1500 tokens but not for 2000 tokens. The vocabulary is thus somewhat smaller for small conversational texts compared to

smaller for the highest token count. These values are therefore only indicative of any effects, and could change when more texts are included.

Table 3.14. Parameter values of Zipf-Mandelbrots law and Zipf's *β*-law for fixed text lengths, per text

		1000 tokens			**1500 tokens**			**2000 tokens**		
	File	**ZM-*α***	**ZM-*β***	**Z-*β***	**ZM-*α***	**ZM-*β***	**Z-*β***	**ZM-*α***	**ZM-*β***	**Z-*β***
1	fn000251	1,112	6,697	1,720	1,191	6,926	1,640			
	fn000260	1,173	6,416	1,744	1,242	7,605	1,681	1,290	8,525	1,671
	fn000279	1,083	6,908	1,716	1,137	7,262	1,632	1,207	8,790	1,637
	fn000284	0,937	1,697	1,733	1,016	2,861	1,692	1,092	4,500	1,702
	fn000300	1,049	3,962	1,663	1,094	4,572	1,668			
	fn000343	1,081	5,608	1,841						
2	fn007056	0,847	0,317	1,953	0,888	0,348	1,864			
	fn007440	0,911	2,306	1,826	0,958	2,460	1,762	1,017	3,488	1,741
	fn007484	1,027	4,554	1,876						
	fn008691	0,880	1,874	1,841						
	fn008781	0,837	1,246	1,864						
	fn008958	1,047	6,160	1,591						
3	fn007147	1,013	3,950	1,959						
	fn007361	1,004	4,503	1,935	1,077	4,873	1,925			
	fn007367	0,954	2,833	1,839	1,004	3,191	1,846	1,054	3,907	1,785
	fn007494	1,034	5,201	1,803	1,066	5,523	1,835	1,133	6,972	1,840
	fn007504	1,038	4,560	1,822	1,033	3,918	1,840	1,079	4,599	1,783
	fn007568	1,189	4,801	1,684	1,175	4,511	1,737			
4	fn000057	0,905	0,403	1,773	0,955	0,686	1,718	0,982	0,784	1,681
	fn000064	0,898	2,216	1,961	0,944	2,595	1,991	0,971	3,151	1,930
	fn000065	1,142	7,712	1,739	1,202	8,591	1,766	1,240	8,299	1,758
	fn000068	0,886	1,157	1,803	0,945	1,339	1,789	0,993	1,676	1,777
	fn000080	1,033	4,100	1,864	1,049	4,015	1,915	1,063	3,465	1,908
	fn000271	0,830	0,731	1,899						
5	Aïsha	0,849	2,437	2,056						
	Dina	0,888	3,108	1,870						
	Elja_Daae	0,834	1,086	2,044						
	Hester	0,809	1,096	2,297						
	Sarra	0,975	4,094	1,882						
	Thamar_Kempees	0,855	2,006	2,052						
6	Belgers	0,780	0,463	2,090	0,850	1,149	2,044			
	Cukier_Leijendekker	0,710	-0,220	2,353						
	Duin_Keultjes	0,779	0,674	2,281						
	Stelling	0,709	0,082	2,366	0,762	0,371	2,283	0,787	0,377	2,270
	Verhagen	0,704	-0,351	2,343	0,750	-0,187	2,307	0,785	-0,003	2,307
	Volkers_Redactie	0,806	0,304	2,182	0,874	0,784	1,992			
7	Bordewijk	0,858	1,637	2,209	0,885	1,687	2,061	0,906	1,632	2,131
	Hildebrand	0,775	0,542	2,391	0,821	0,845	2,384	0,853	1,135	2,332
	Multatuli	0,802	1,137	2,128	0,877	2,327	2,094	0,929	3,115	1,990
	Nieuwenhuis	0,849	2,158	2,213	0,913	3,087	2,038	0,943	3,288	2,048
	Van_den_Vondel	0,704	-0,036	2,408	0,751	0,211	2,339	0,806	0,727	2,280
	Van_Eeden	0,859	1,813	2,174	0,879	1,519	2,122	0,897	1,363	2,107

Table 3.15. Mean parameter values of Zipf-Mandelbrots law and Zipf's β-law for fixed text lengths per category

	No. of texts	**1000 tokens**			**No. of texts**	**1500 tokens**			**No. of texts**	**2000 tokens**		
		ZM-α	**ZM-β**	**Z-β**		**ZM-α**	**ZM-β**	**Z-β**		**ZM-α**	**ZM-β**	**Z-β**
1. Conversations	6	1,072	5,215	1,736	5	1,136	5,845	1,663	3	1,196	7,272	1,670
2. Commentaries	6	0,925	2,743	1,825	2	0,923	1,404	1,813	1	1,017	3,488	1,741
3. Discussions	6	1,039	4,308	1,840	5	1,071	4,403	1,837	3	1,089	5,159	1,802
4. Speeches	6	0,949	2,720	1,840	5	1,019	3,445	1,836	5	1,050	3,475	1,811
5. Blogs	6	0,868	2,305	2,034	0	NA	NA	NA	0	NA	NA	NA
6. Articles	6	0,748	0,159	2,269	4	0,809	0,529	2,157	2	0,786	0,187	2,288
7. Literature	6	0,808	1,209	2,254	6	0,855	1,613	2,173	6	0,889	1,877	2,148

discussions, but this difference disappears when longer texts are considered. Between the groups of spoken and written categories, conversations and discussions display a higher ZM-α than all written categories for all token counts. In all cases, conversations and discussions display less varied vocabularies than written texts. Commentaries have a higher ZM-α than articles for 1000 tokens, but not for 1500 or 2000 tokens. This difference thus does not seem to be meaningful. Speeches have a higher ZM-α (thus less varied vocabulary) than articles for all token counts, and a higher ZM-α than literature for 1500 and 2000 tokens. For p-values, see Table 3.16.

Also for ZM-β-values per category, it is found that there are significant differences for all token counts, as shown by a one-way analysis of variance (1000 tokens: F(6,35) = 6,332, p < 0,001; 1500 tokens: F(5,21) = 5,693, p = 0,002; 2000 tokens: F(5,14) = 4,406, p = 0,013). A Tukey's HSD post-hoc test reveals that no differences exist within the group of written categories or within the group of spoken categories. Between the spoken and written category, it is found that conversations display a higher ZM-β than articles and literature for all token counts, meaning that the curvature for the highest ranks of Zipf-Mandelbrots law is more pronounced for this category. Discussions display a higher ZM-β than articles and literature for 1000 tokens only. No other differences are found. For p-values, see Table 3.16.

Significant differences are found for Z-β-values per category for all token counts as well, as shown by a one-way analysis of variance (1000 tokens: F(6,35) = 23,248, p < 0,001; 1500 tokens: F(5,21) = 16,233, p < 0,001; 2000 tokens: F(5,14) = 18,324, p < 0,001). A Tukey's HSD post-hoc test reveals that no differences exist within the category of spoken texts. Within the category of written texts, blogs display a higher Z-β than articles and literature for 1000 tokens, reflecting the smaller vocabulary of blogs. No blog reached 1500 tokens so it could not be tested if these differences still exist when text size increases. Between written and spoken texts, all differences are significant except for the difference between discussions and blogs and between speeches and blogs for 1000 tokens (which both approached significance with p = 0,06), with written texts displaying a higher Z-β than spoken texts. Save these exceptions, spoken texts have smaller vocabularies than written texts. For p-values, see Table 3.16.

3.2.3 Discussion

Texts on a continuum from spontaneously spoken text to extensively thought through written text were examined to investigate the influence of modality and text length on the parameters of Zipf's law. Growth curves were used to see if parameter values develop in the same way for different kinds of texts, and to find out when the change in the parameter values stabilizes.

It was found that the growth curves for ZM-α behave the same for all texts. In all cases, ZM-α quickly increases for very small samples, after which the increase levels off and continues at a slow but seemingly constant rate. The shape of the growth curve could accurately be described with a logarithmic function.

The current study only concerns small sample sizes. The largest samples were 5000 tokens, which, in terms of Zipf's law, is still rather small. It is therefore based on the current data set unknown how the growth curves will develop after this number of tokens. If the curve continues to follow a logarithmic curve then that means that ZM-α continues to grow – theoretically – indefinitely. But that is incompatible with the omnipresent idea that ZM-$\alpha \approx 1$. How can these two things be combined?

The mean parameter values of all ZM-α growth curves were A = 0,166 and B = -0,24. This means that the average ZM-α growth curve can be described as

9) $\text{ZM-}\alpha(x) = 0{,}166 \cdot \log x - 0{,}24$

were x is the number of tokens in the text. Say, ZM-$\alpha \approx 1$ concerns the range $0{,}7 \leq \text{ZM-}\alpha \leq 1{,}3$. Solving the equation in 9 for these values shows that this range of ZM-α values is found for texts of size $288 \leq x \leq 10.690$. So, if it is true that the growth of ZM-α continues at exactly the same rate as given by the growth curves fitted here, then a value of which it can be said that it is approximately 1 can be fitted to text sizes up to 10.000 tokens. This is where the idea of a constant comes from: there is none, but it seems like it because ZM-α values change little for such a wide range of number of tokens.

Table 3.16. p-values for ZM-*α*, ZM-*β* and Z-*β* for 1000, 1500 and 2000 tokens

		1000 tokens						**1500 tokens**						**2000 tokens**					
		7. literature	**6. articles**	**5. blogs**	**4. speeches**	**3. discussions**	**2. commentaries**	**7. literature**	**6. articles**	**5. blogs**	**4. speeches**	**3. discussions**	**2. commentaries**	**7. literature**	**6. articles**	**5. blogs**	**4. speeches**	**3. discussions**	**2. commentaries**
ZM-*α*	**1. conversations**	*0,000*	*0,000*	*0,001*	0,123	0,989	0,037	*0,000*	*0,000*	NA	0,212	0,773	0,038	*0,001*	*0,001*	NA	0,174	0,566	0,401
	2. commentaries	0,165	*0,001*	0,872	0,998	0,188		0,885	0,556	NA	0,687	0,252		0,665	0,221	NA	0,999	0,965	
	3. discussions	*0,000*	*0,000*	0,010	0,448			*0,002*	*0,001*	NA	0,894			0,028	0,009	NA	0,982		
	4. speeches	0,052	*0,002*	0,572				0,024	0,007	NA				0,042	0,013	NA			
	5. blogs	0,832	0,142					NA	NA					NA	NA				
	6. articles	0,839						0,942						0,608					
ZM-*β*	**1. conversations**	*0,004*	*0,000*	0,065	0,163	0,963	0,170	0,010	*0,003*	NA	0,321	0,799	0,072	0,019	0,017	NA	0,164	0,786	0,592
	2. commentaries	0,690	0,135	0,999	1,000	0,671		1,000	0,993	NA	0,752	0,379		0,973	0,758	NA	1,000	0,976	
	3. discussions	0,041	*0,002*	0,390	0,656			0,152	0,423	NA	0,956			0,251	0,134	NA	0,853		
	4. speeches	0,705	0,142	0,999				0,558	0,196	NA				0,772	0,410	NA			
	5. blogs	0,913	0,311					NA	NA					NA	NA				
	6. articles	0,928						0,933						0,900					
Z-*β*	**1. conversations**	*0,000*	*0,000*	*0,001*	0,667	0,662	0,800	*0,000*	*0,000*	NA	0,192	0,188	0,611	*0,000*	*0,000*	NA	0,417	0,588	0,988
	2. commentaries	*0,000*	*0,000*	*0,036*	1,000	1,000		0,009	0,022	NA	1,000	1,000		0,019	0,005	NA	0,985	0,993	
	3. discussions	*0,000*	*0,000*	0,063	1,000			*0,001*	*0,004*	NA	1,000			*0,002*	*0,001*	NA	1,000		
	4. speeches	*0,000*	*0,000*	0,061				*0,001*	*0,004*	NA				*0,001*	*0,001*	NA			
	5. blogs	0,022	0,012					NA	NA					NA	NA				
	6. articles	1,000						1,000						0,532					

An upper limit to number of tokens also exists for another reason. Theoretically, new words can be made up and thus added to a text at any time, at least for Dutch and most other Western languages. In written texts, typos occur, which also become part of the hapax legomena. But there comes a time when the normal vocabulary is exhausted, and long before that time it will become a rare event to encounter new tokens. This is the point where what Piotrowski and Spivak (1994) call *saturation* starts to occur. Consider a growing text. If no new types are added, then every newly encountered token has already been used before. Due to the logarithmic scaling, the effect of this is large on low frequency items and small on high frequency items. An already frequent word that is encountered again stays frequent and is unlikely to move to a different rank. A low frequency item, however, especially one that previously occurred only once or twice, will get a different rank when it is encountered again. If this happens often while no new tokens are encountered then the result is that there are many more types in the mid-frequency range, and fewer types in the low frequency range. The result is a curving distribution. The point at which this happens is highly dependent on vocabulary size. Most texts never reach this number. This is different for Modern Simplified Chinese. For this language, the number of characters is fixed at 8105. Dahui, Menghui and Zengru (2005) found a curved distribution for Modern Simplified Chinese (7 texts, ranging from 142.549 to 2.835.949 tokens) (also discussed in Chapter 2, Section 2.2.1). They found that mid frequencies were overrepresented, while the number of low frequency items was much smaller than that usually found in Zipf's law. Western languages will reach this point much later, but it will be reached eventually.

For ZM-β, no pattern could be discovered in the shape of the growth curves. Values cover a wide range, going from −0,8 to 10,5. For some texts ZM-β remains relatively stable as text size increases, for others it increases, and in some cases, it decreases. Some, but not all, seem to display two-phase behaviour with ZM-β growing linearly until it levels of quite abruptly. To be able to interpret this finding, it is important to remember what ZM-β is for. As discussed in Chapter 2, it was introduced by Benoit Mandelbrot to obtain a better fit of Zipf's law for the highest rank numbers (Mandelbrot, 1954). A positive number means that the slope of Zipf's law predicted by the ZM-α-value is an overestimation for the most highly ranked tokens (meaning that high frequency items are less frequent than expected based on the slope); a negative number means that this slope is an underestimation (meaning that high frequency items are even more frequent than expected based on the slope). In all cases, ZM-β is calculated based on only a handful of types and thus data points, especially with small sized samples as are used here. This renders the value of ZM-β prone to fluctuations, which explains the wide range of values and variation in growth curves found here. The first few ranks are always occupied by function words. These words mostly serve syntactical purposes, rather than contributing to the content of the message. This suggests that the usage of function words varies greatly between texts. Qualitative analyses of the texts are required to gain more insight into these workings.

These large fluctuations do not mean that the parameter is useless. On the contrary: including ZM-β in the formula of Zipf's law assures a more accurate calculation of ZM-α: thanks to ZM-β, ZM-α is less influenced by the deviation from a straight line seen for the first few ranks. What it does mean is that ZM-β is rather useless as a characterization of a text and should not be used as such. It is not possible to use ZM-β for comparisons amongst texts.

The results for Z-β were unexpected. Visual inspection of the growth curves suggested that they were all rather similar: Z-β either decreases slightly when text size grows or stays relatively stable. Nevertheless, no pattern could be discovered when comparing them statistically. Excluding the first few data points – for which fluctuation seemed to be most pronounced – did not improve the results. It is expected that a pattern can be found when larger text sizes are analysed, which was sadly not possible with the current sources. However, although it can be concluded that no general pattern of growth can be discovered, it is still the case that values of Z-β stay relatively stable as text size increases. They do not largely fluctuate. This means that despite the absence of uniform growth curves, it can still be considered meaningful to compare the values of Z-β for texts with identical token counts of different sources. A higher Z-β reflects a more varied vocabulary than a low Z-β: it means that the first few frequency classes, the words with frequency 1 or 2, are proportionally larger than the high frequency classes.

The growth curves for ZM-α and for Z-β were used to determine the minimal workable text size, the text size at which the rate with which the value of the parameter changes stabilizes. It turns out that this happens around 1500 tokens. To be on the safe side it is advised to analyse at least 2000 tokens when studying Zipf's law, whenever possible.

Now that the minimal workable text size is known, texts of different sources could be compared to see if and how they differ. As a baseline, type/token ratios were compared, which revealed that spoken texts have larger text-size independent TTR's than written texts. Blogs and commentaries also do not differ.

The downside of TTR's is that it is only one value that is being used as an index of vocabulary richness. Aberrations for high frequency items only, for example, cannot be detected. A truly aberrant frequency distribution is not detected either. Zipf's law offers tools for that. The value of ZM-β indicates if the frequencies of the highest frequency items are more (in the case of a high ZM-β) or less uniform (in the case of a negative ZM-*β*) than in the rest of the text. The values of ZM-*α* and Z-*β* reflect vocabulary richness in the rest of the text. Zipf-Mandelbrots law can be calculated for somewhat smaller text samples than Zipf´s *β*-law (very small samples result in empty frequency classes), but Zipf’s *β*-law is more sensitive to disruptions because of the absence of ranking. Zipf's law can thus provide a valuable view on word frequencies in a text, even detached from the theoretical framework discussed in Chapter 2.

Unfortunately, not all texts in the current study reached the advised 1500 tokens for studying Zipf's law. Longer samples were sadly not available from the current sources. The values of the parameters were therefore compared for 1000, 1500 and 2000 tokens to see if any pattern arises. Clear differences were found between written and spoken texts, with blogs clearly on a boundary position. This difference is most clearly reflected by Z-*β*, and to a lesser extent by ZM-*α*.

ZM-*β* did not reflect the difference between written and spoken texts very well. Given the large fluctuations in its growth curves, this is no unexpected result. No differences were found within the groups of written and spoken categories, but also between groups most differences did not reach significance. Only the difference between conversations and articles and literature remained, the text categories that are furthest apart on the continuum of categories that was assumed. This suggests that there might be a difference between categories, but the current sources were not sufficient to properly investigate this difference.

As expected, parameter values of ZM-*α* and Z-*β* reflected a larger vocabulary in written texts than in spoken texts, even if the spoken texts were prepared beforehand (e.g. sermons/speeches). ZM-*α* was higher for spoken texts than for written texts, Z-*β* was lower. This is in line with Ridley (1982)’s results for Zipf's law in spoken texts. The word frequency distributions thus suggest that the writers of such texts take into account that their text is going to be spoken aloud. A larger vocabulary is thus not a sign of the amount to which the text was prepared. Rather, it seems to be the case that a smaller vocabulary is necessary to keep the spoken text accessible. This makes sense, considering that the listener cannot pause, slow down or repeat part of the text as a reader can. This is (either consciously or unconsciously) taken into account when a spoken text is prepared.

Zipf's law thus does not behave radically different for written or spoken texts. But parameter values are not exactly the same either. ZM-*α* is generally higher; Z-*β* is lower. This finding confirms the sensibility to treat ZM-*α* and Z-*β* as parameters rather than constants, an approach that is now usually (but still not always, e.g. Yang, 2013) the general practice.

3.3 Part II: Removing Zipf's law

The fact that Zipf's law is found in every text in every language means that – to our current knowledge – there are no naturally occurring texts for which Zipf's law does *not* apply. It is currently unknown if it is at all possible to remove Zipf's law from a text, and how readers would respond to such a text. In this chapter, I will attempt to create such a text, and explore how naïve readers respond to it. A third, otherwise adapted text is used as a control condition.

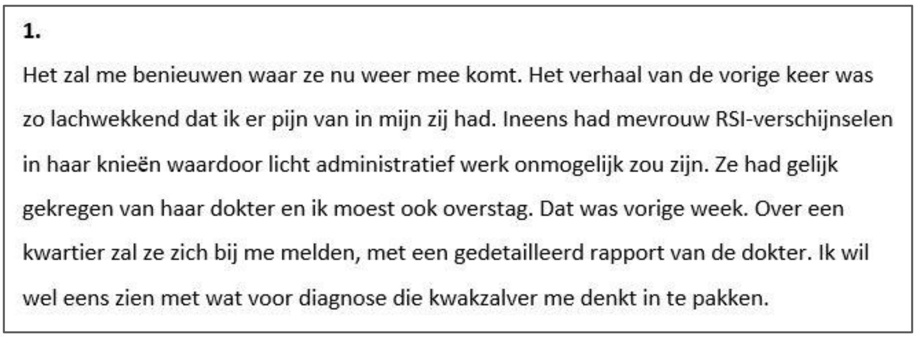
One practical problem is that no definition exists of when Zipf's law does not apply. Good or bad fit in terms of R^2 can be used, but when is a fit considered too bad? When do parameter values fall within the normal range, and when are they too extreme to be considered part of a Zipfian distribution? Above, it became clear that there are many factors that influence the value of the parameters of Zipf's law, of which medium and text length are amongst the most important ones. As a result of this, parameter values can cover quite a wide range of values. In addition, Zipf-Mandelbrot's law makes use of the ranking of values, which means that the frequency distribution necessarily follows a decreasing slope. This means that any disruptions to the distribution have to be quite extreme to be noticeable.
I do not have an answer to this issue. Instead, I take a practical approach. I assume that the slope of Zipf's β-law has to be log-linearly decreasing, with a good fit (compared to a comparable control text). Above, it was already mentioned that this version of Zipf's law is more useful for the detection of disruptions to Zipf's law than Zipf-Mandelbrot's law. Parameter values are compared between comparable texts of equal length. In addition, I visually inspect the frequency distributions in search of disruptions. The question whether or not a frequency distribution conforms to Zipf's law is hard to answer with a clear *yes* or *no*, but rather concerns a continuum. The use of control texts is a practical way to determine the 'normal' values for a comparable text. Large deviations from these values can be considered a sign that Zipf's law might be disrupted.

The method itself that is being applied here is one that is omnipresent in linguistics. Linguists (especially generative linguistics) generally take a word or sentence, modify it in some way, and ask naïve speakers of the language or dialect under investigation to give their judgment (e.g. "Who did Susan ask why Sam was waiting for?"). If the speakers do not accept the word or sentence then the linguist will come up with a linguistic explanation (e.g. violation of island constraints). The current experiment attempts to make use of the same logic, but on a textual level: take a text, modify it in some way, and ask naïve speakers for their opinion. Afterwards, I will speculate about a (linguistic) explanation. This study should be seen as a first exploratory study into this topic.

3.3.1 Methods

The texts

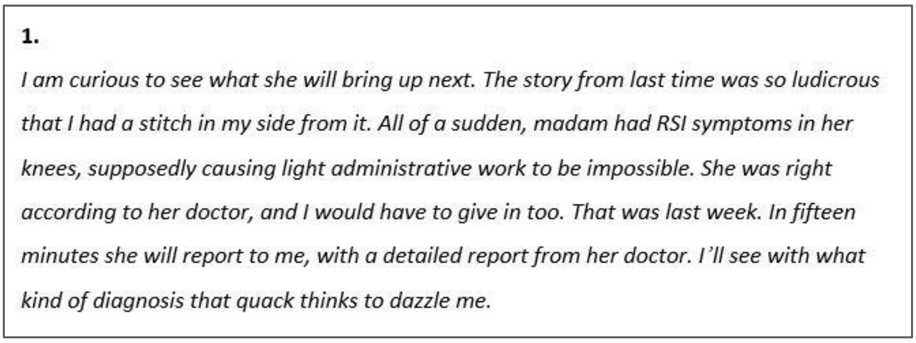
The original text consists of the first 954 words of the Dutch novel *Ik ben omringd door debielen en ik voel me goed* (I am surrounded by morons and I feel good) by Stefan Nieuwenhuis (2005) (at 954 words a natural boundary exists in the text). This text is fully grammatical but written in a marked style. It is written from the point of view of a sarcastic and bitter health and safety officer, who is about to call in a patient. He grumbles about ‘her kind of woman’, the cleaners of his office and the Dutch social care system. A small excerpt is given in Figure 3.8 (all excerpts are translated by me), the full text is given in Appendix A1.



1.

Het zal me benieuwen waar ze nu weer mee komt. Het verhaal van de vorige keer was zo lachwekkend dat ik er pijn van in mijn zij had. Ineens had mevrouw RSI-verschijnselen in haar knieën waardoor licht administratief werk onmogelijk zou zijn. Ze had gelijk gekregen van haar dokter en ik moest ook overstag. Dat was vorige week. Over een kwartier zal ze zich bij me melden, met een gedetailleerd rapport van de dokter. Ik wil wel eens zien met wat voor diagnose die kwakzalver me denkt in te pakken.

Figure 3.8a. Excerpt of the original text



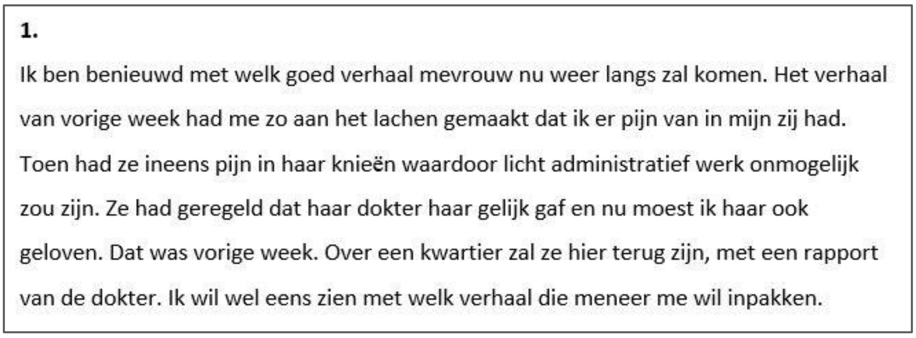
1.

I am curious to see what she will bring up next. The story from last time was so ludicrous that I had a stitch in my side from it. All of a sudden, madam had RSI symptoms in her knees, supposedly causing light administrative work to be impossible. She was right according to her doctor, and I would have to give in too. That was last week. In fifteen minutes she will report to me, with a detailed report from her doctor. I’ll see with what kind of diagnosis that quack thinks to dazzle me.

Figure 3.8b. English translation of the excerpt of the original text

Removing Zipf's law means changing word frequencies. This can be done in different ways. A gap in the frequency distribution can be created by having only very high and very low frequencies, or they can follow a uniform distribution by all having almost the same frequency. The latter is the approach taken here. One problem in this is that the text had to remain readable and conform to the syntactic rules of Dutch. It was therefore decided to only change word frequencies of lexical words, not of the grammatical words. A small excerpt of the resulting text is given in Figure 3.9, the full text can be found in Appendix A2.

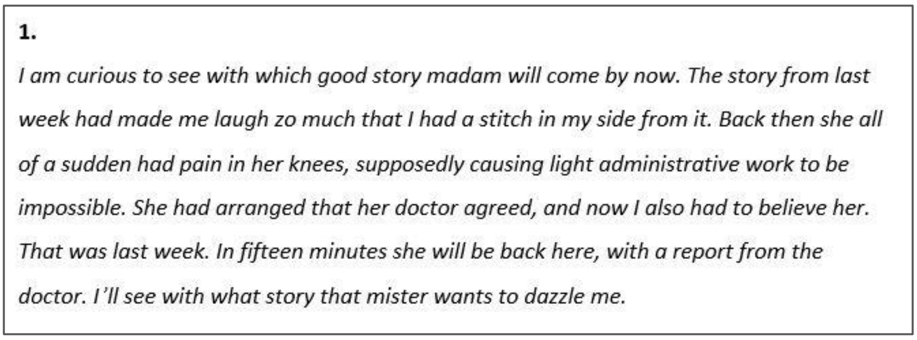
A third text was needed to examine the effect of the manipulations of Zipf's law compared to manipulations in general. This text should be a modified version of the original, but with a word frequency distribution that is identical to the original. This text was obtained by changing words for their (near) synonyms, changing some long sentences into short ones and vice versa, and by replacing non-stressed pronouns ('ze', 'me') by their stressed counterparts ('zij', 'mij'). An excerpt of the resulting text is given in Figure 3.10, the full text can be found in Appendix A3.



1.

Ik ben benieuwd met welk goed verhaal mevrouw nu weer langs zal komen. Het verhaal van vorige week had me zo aan het lachen gemaakt dat ik er pijn van in mijn zij had. Toen had ze ineens pijn in haar knieën waardoor licht administratief werk onmogelijk zou zijn. Ze had geregeld dat haar dokter haar gelijk gaf en nu moest ik haar ook geloven. Dat was vorige week. Over een kwartier zal ze hier terug zijn, met een rapport van de dokter. Ik wil wel eens zien met welk verhaal die meneer me wil inpakken.

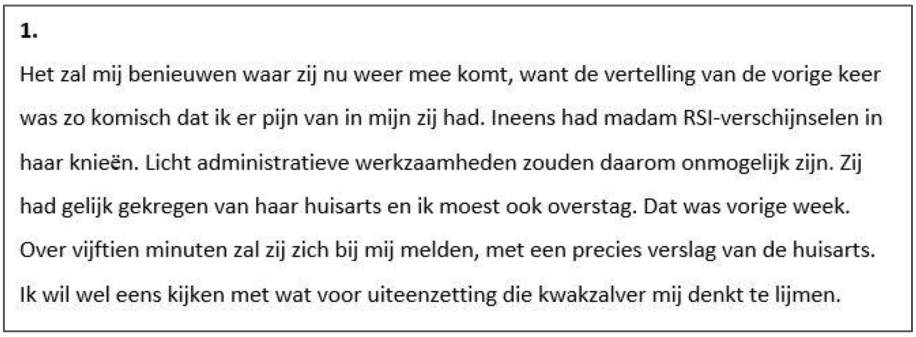
Figure 3.9a. Excerpt of the text without Zipf's law



1.

I am curious to see with which good story madam will come by now. The story from last week had made me laugh zo much that I had a stitch in my side from it. Back then she all of a sudden had pain in her knees, supposedly causing light administrative work to be impossible. She had arranged that her doctor agreed, and now I also had to believe her. That was last week. In fifteen minutes she will be back here, with a report from the doctor. I'll see with what story that mister wants to dazzle me.

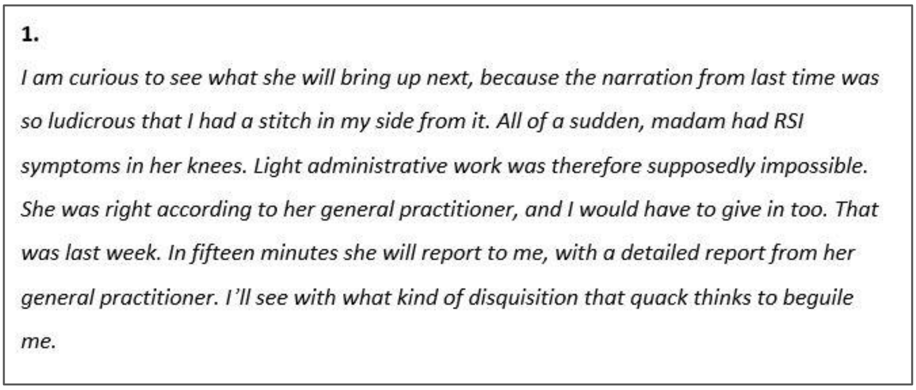
Figure 3.9b. English translation of the excerpt of the text without Zipf's law



1.

Het zal mij benieuwen waar zij nu weer mee komt, want de vertelling van de vorige keer was zo komisch dat ik er pijn van in mijn zij had. Ineens had madam RSI-verschijnselen in haar knieën. Licht administratieve werkzaamheden zouden daarom onmogelijk zijn. Zij had gelijk gekregen van haar huisarts en ik moest ook overstag. Dat was vorige week. Over vijftien minuten zal zij zich bij mij melden, met een precies verslag van de huisarts. Ik wil wel eens kijken met wat voor uiteenzetting die kwakzalver mij denkt te lijmen.

Figure 3.10a. Control text



1.

I am curious to see what she will bring up next, because the narration from last time was so ludicrous that I had a stitch in my side from it. All of a sudden, madam had RSI symptoms in her knees. Light administrative work was therefore supposedly impossible. She was right according to her general practitioner, and I would have to give in too. That was last week. In fifteen minutes she will report to me, with a detailed report from her general practitioner. I'll see with what kind of disquisition that quack thinks to beguile me.

Figure 3.10b. English translation of the excerpt of the control text

Text characteristics

The three texts hardly differ in terms of number of sentences or number of clauses. The TTR is slightly higher for the text without Zipf's law, which is a direct result of the way in which the text was altered. See Table 3.17 for the exact values.

As elsewhere, punctuation and capitals were removed for the fitting of Zipf's law. Accents used to indicate stress were ignored. The parameters of Zipf's law were then calculated for all words and for lexical words and grammatical words separately. It should be kept in mind that the length of the text was such that these values have no absolute meaning, as it follows from Part I of this chapter that text length should at least be 1500-2000 tokens. In this case, the parameters of Zipf's law were not calculated to compare the texts, but to see if they conform to Zipf's law in the first

Table 3.17. Text characteristics

	$\textbf{TTR}_{\textbf{dep}}$	$\textbf{TTR}_{\textbf{indep}}$	**Number of sentences**	**Number of clauses**	**% Hapax legomena**
Original	0,468	0,550	77	137	71,9
No Zipf	0,327	0,520	77	140	31,6
Control	0,486	0,567	75	142	74,3

place. All parameter values are given in Table 3.18. The distributions of Zipf-Mandelbrots law and Zipf's β-law are given in Figure 3.11 and 3.12.

The parameters for Zipf-Mandelbrots law for the full text adapted to not conform to Zipf's law are actually not too different from those found for the original and control text: ZM-α = 0,81 and ZM-β = 2,78, while the values for the original are ZM-α = 0,85 and ZM-β = 2,15 and for the control text ZM-α = 0,82 and ZM-β = 1,75. The ZM-R^2 of the non-Zipfian text is as high as ZM-R^2 = 0,99. This means that the formula of Zipf-Mandelbrots law can accurately describe the distribution of the word frequencies of this text when all words are considered. Apparently, it is the case that the normal Zipfian distribution of the grammatical words conceals the deviations of the distribution of the lexical words. Nevertheless, there are some visible differences compared the normal text, especially concerning the lower frequencies. Usually, it is found that the words with frequency 1, the hapax legomena, make up about 50% of the text. This percentage is often even larger for shorter texts. Here, the original text has 71,9% hapax legomena; the control text has 74,0 % hapax legomena. Because of the logarithmic scaling, the hapax legomena show up in the plots as a long vertical line at the bottom of the distribution. For the non-Zipfian text this line is visibly much shorter. The percentage of hapax legomena is as low as 31,6%. Although not captured by the formula of Zipf-Mandelbrots law, the percentage of hapax legomena shows that the word frequency distribution of the non-Zipfian text is exceptional.

The parameter values for the content words in the text without Zipf's law are clearly out of the ordinary. The value that is found for ZM-β is as high as 204,42, while ZM-β for the lexical words in the other texts is -0,52 for the original text and -0,65 for the control text. ZM-β captures the curvature of the distribution for the highest ranks. This very high value thus means that this curvature is much more pronounced than for the other two texts, for which it is almost absent and if anything, curving in the opposite direction. This difference is also clearly seen in Figure 3.11. This high ZM-β means that the slope given by ZM-α is inaccurate not only for the first few ranks, but for almost the entire distribution (since the highest rank number in this distribution is 200). This behaviour of the parameters is so different from that of the other texts that it seems safe to conclude that the lexical words in this text indeed do not conform to Zipf's law.

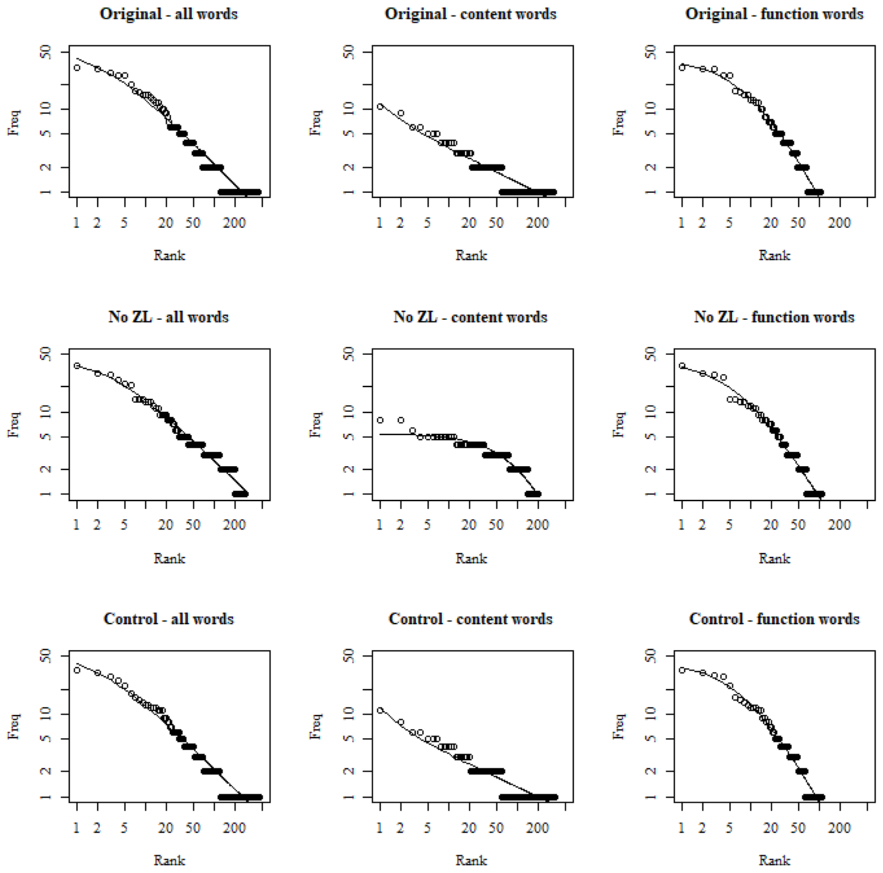


Figure 3.11. ZM-α for the original text, the text adapted to not conform to Zipf's law and the control text, adapted but still conforming to Zipf's law, for all words and for lexical words and grammatical words separately.

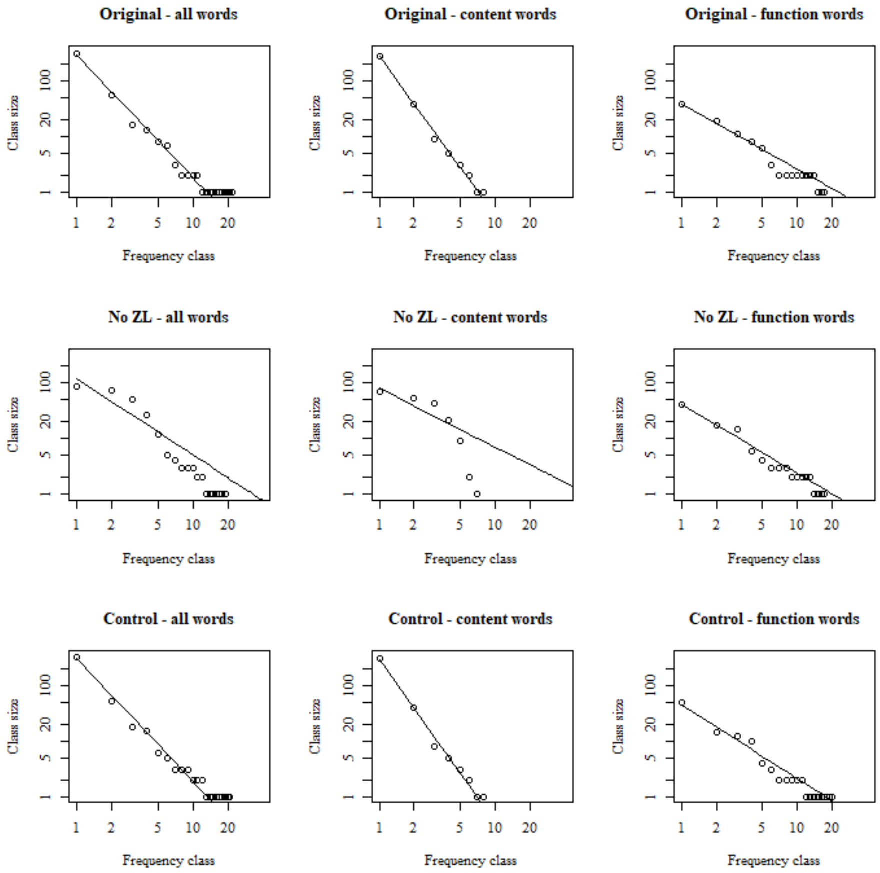


Figure 3.12. Z-β in the original text, the text adapted to not conform to Zipf's law and the control text, adapted but still conforming to Zipf's law, for all words and for lexical words and grammatical words separately.

Table 3.18. Parameter values for the original text, the text adapted to not conform to Zipf's law and the control text, adapted but still conforming to Zipf's law.

Text	Analysis	Tokens	Types	ZM-α	ZM-β	ZM-R^2	Z-β	Z-R^2
Original	All words	960	428	0,851	2,153	0,964	2,203	0,995
	Lexical words	455	343	0,413	-0,519	0,933	2,858	1,000
	Gramm. words	503	105	1,611	9,676	0,984	1,170	0,975
No Zipf	All words	940	277	0,813	2,771	0,985	1,363	0,782
	Lexical words	473	200	2,530	204,420	0,899	1,059	0,526
	Gramm. words	466	107	1,366	6,410	0,978	1,233	0,982
Control	All words	958	441	0,821	1,747	0,972	2,217	0,993
	Lexical words	465	356	0,393	-0,648	0,926	2,933	1,000
	Gramm. words	493	110	1,535	8,357	0,987	1,333	0,979

In the discussion to Part I of this chapter, it was already mentioned that Zipf's β-law is more sensitive to disruptions because of the absence of ranking. This also becomes clear here: the parameter Z-β = 1,36 that is found for the full non-Zipfian text is actually not too far away from the values Z-β = 2,20 for the original text and Z-β = 2,22 for the control text. However, the fit of the formula Z-R^2 = 0,78 is much lower, as usually values above 0,9 are found. The value for the fit of the lexical words Z-R^2 = 0,53 is (unsurprisingly) even lower than that of the analysis including all words. The plots clearly show that the frequency class distribution of the non-Zipfian text is curved instead of log-linear.

It seems safe to conclude that I succeeded at constructing a text in which the content words do not conform to Zipf's law. Zipf's law for all words is disrupted, but possibly not altogether absent.

Questionnaire

A questionnaire was constructed to test how people respond to the different texts. The goal of this questionnaire was to help respondents give the right kind of feedback. There were no expectations beforehand as to what kind of responses readers would give, so the questionnaire was constructed to cover a wide range of text phenomena.

All texts were introduced to participants in the same way, namely as a text translated from English to Dutch by a beginning translator. This introduction was chosen because it suggests that something might be wrong with the text, but no hints are given about the kind of thing that might be wrong with the text.

The questionnaire consisted of the following questions:

A 1st question: What is your first opinion? (open question)
B 41 text rating questions (7-point Likert scale)
C 3 questions about the variation in verbs, nouns, and adverbs/adjectives
D Four questions about errors: Does this text contain … kind of errors?
E Text rating questions: Rate the text on a scale from 1-10
F Exclusion and control questions

The full questionnaire is given in Appendix B.

Participants

Three different groups participated in this study. Each group read one of the texts. All participants were university students. The original text was given to 32 participants (5 male, mean age 20.0 yrs.). One participant was excluded because of incomplete answers. The non-Zipfian text was handed to 57 participants (17 male, mean age 21.0 yrs.). Two participants were excluded because of incomplete questionnaires. The control text was handed to 28 participants (5 male, mean age 21 yrs.).

Analysis

Results from the questions in part B, C and D concern ordinal data. For part D, results were converted to numeric outcomes: no = 0; yes, some = 1; yes, many = 2. Kruskal–Wallis tests by rank are used to test for any differences between groups, followed by Dunn's (1964) test of multiple comparisons in case of a significant difference between groups. The ratings given in Part E are compared using a one-way analysis of variance, followed up by a Tukey's HSD post-hoc test.

3.3.2 Results

Questionnaire

Question 53 in Part F ("Did you read this text or a text very similar to this text ever before?") was included as an exclusion criterion. Based on this question, two participants in the no Zipf's law-group (both m, 19 yrs.) and one participant in the control group (m, 20 yrs.) were excluded, because they filled in that they had read something very similar.

Part A | Question 1: First opinion

The first question asked for a first opinion. All answers given to this question are provided in Appendix C. A few points were made remarkably often. These points are given in Table 3.19. A comment about many word repetitions was made by 12% of the participants in the No Zipf-group, while none of the participants in the other group made a comment about that. The manipulations were thus noticed. Striking use of pronouns, on the other hand, was mentioned by 29% of the participants in the control group but no one in the other groups. Again, this is a direct result of the applied manipulations. Comments about aberrant, weird, archaic or wrong choice of words were made by participants in all three groups, but strikingly more so by participants in the control group (original: 27%; no Zipf's law: 26%; control: 64%). This is thus a feature of all texts but exaggerated in the control text as a result of the applied manipulations.

The task description was of influence as well. Participants in all groups claimed to see influence of the (in fact non-existent) English source text. Control participants made this comment most often (original: 10%; no Zipf’s law: 14%; control: 43%). Problems with reading fluency were reported by participants in all groups, but most in the no-Zipf's law group (original: 10%; no Zipf's law: 26%; control: 2%). Also mentioned in all groups were strikingly many short sentences (original: 11%; no Zipf's law: 25%; control: 13%), and weird sentence constructions (original: 18%; no Zipf's law: 28%; control: 23%). About 19% of the participants in the original group thought the text was easy, a thought shared by 23% of the no Zipf's law group. None of the control participants mentioned this.

Table 3.19. Points made remarkably often in response to the question for a first opinion

Point	Original	No Zipf	Control
Many word repetitions	0%	12%	0%
Striking use of pronouns	0%	0%	29%
Aberrant/weird/archaic/wrong words	27%	26%	64%
Influence of the English source text	10%	14%	43%
Problems with reading fluency	10%	26%	2%
Many short sentences	11%	25%	13%
Weird sentence constructions	18%	28%	23%
Many easy words / easy text	19%	23%	0%

Part B | Question 2 – 42: Text rating questions

Mean values and *p*-values for all questions for which significant differences were found are given in Table 3.20.

Table 3.20. Mean values, medians, *p*- and Z-values for the text rating questions for which significant differences were found.

	Original		No ZL		Control		Kruskal		Z-values post hoc test			values post hoc test		
Q	mean	median	mean	median	mean	median	X^2	p	cont-no ZL	or-no ZL	or-cont	cont-no ZL	or-no ZL	or-cont
3	2,613	2	1,929	2	3,036	3	15,606	0,000	-3,682	-2,614	1,024	*0,001*	*0,013*	0,306
5	4,323	4	3,386	3	2,964	3	10,793	0,005	1,126	-2,518	-3,152	0,260	*0,018*	*0,005*
6	2,710	2	3,386	3	3,857	3,5	6,709	0,035	-1,231	1,723	2,565	0,218	0,127	*0,031*
7	3,355	3	4,746	5	5,071	5	16,826	0,000	-0,767	3,528	3,698	0,443	*0,001*	*0,001*
8	3,935	4	4,912	5	5,464	6	14,531	0,001	-1,560	2,739	3,725	0,119	*0,009*	*0,001*
11	3,645	3	4,684	5	4,250	5	8,458	0,015	1,396	2,874	1,225	0,244	*0,012*	0,221
12	4,258	5	4,474	5	6,036	6	23,974	0,000	-4,391	0,541	4,350	*0,000*	0,589	*0,000*
14	4,161	4	3,702	3	4,607	5	6,157	0,046	-2,459	-1,154	1,189	*0,042*	0,249	0,352
17	4,452	4	3,684	3	3,393	3	6,579	0,037	0,709	-2,076	-2,404	0,479	0,057	*0,049*
18	3,452	3	4,193	5	3,929	4	6,937	0,031	0,885	2,634	1,471	0,376	*0,025*	0,212
24	3,516	3	4,421	5	4,821	5	9,016	0,011	-0,865	2,409	2,827	0,387	*0,024*	*0,014*
26	3,968	4	4,965	5	5,357	6	12,337	0,002	-0,890	2,891	3,263	0,373	*0,006*	*0,003*
27	4,233	4,5	4,509	5	5,500	6	11,376	0,003	-2,873	0,696	3,120	*0,006*	0,487	*0,005*
28	4,129	4	4,875	5	4,857	5	6,349	0,042	0,307	2,454	1,834	0,759	*0,042*	0,100
29	4,516	5	4,684	5	6,143	7	18,497	0,000	-3,780	0,641	3,895	*0,000*	0,521	*0,000*
31	2,839	3	2,895	3	3,704	4	6,062	0,048	-2,307	0,032	2,075	0,063	0,974	0,057
34	2,935	3	3,158	3	4,000	4	8,315	0,016	-2,438	0,617	2,686	*0,022*	0,537	*0,022*
36	2,871	2	4,088	5	4,250	5	10,886	0,004	-0,419	2,934	2,882	0,675	*0,010*	*0,006*
39	3,774	3	4,772	5	5,036	5	11,833	0,003	-0,938	2,792	3,220	0,348	*0,008*	*0,004*
40	4,129	4	3,404	3	3,607	3	6,749	0,034	-0,518	-2,570	-1,741	0,605	*0,031*	0,122

The questions on which the original text is rated differently than the two adapted texts are the following:

5. This text is pleasant to read (*original participants agree more*)
7. The writing style of this text is unpleasant (*original participants agree less*)
8. The text feels unnatural and constructed (*original participants agree less*).
24. This text is cumbersome to read (*original participants agree less*)
26. The text is fuzzy (*original participants agree less*).
36. The text is difficult to read (*original participants agree less*).
39. The rhythm of the text is unnatural (*original participants agree less*).

Generally speaking, the original text is thus rated as more natural, and more pleasant and easy to read than the two adapted texts. On only one questions was the text without Zipf's law rated differently than the original text and the control text:

3. This text is poetic (*no ZL-participants agree less*)

The control text is rated differently than the original text and the text without Zipf's law on the following questions:

12. This text contains aberrant choice of words (*control participants agree more*)
27. This text contains strangely constructed sentences (*control participants agree more*)
29. The author of this text used words in a way in which I normally would not use these words (*control participants agree more*)
31. This text is marked by official language use (*control participants agree more*)
34. This text contains old-fashioned language use (*control participants agree more*)

These results directly reflect the manipulations that were applied.

Questions on which the original text is rated differently than the text without Zipf's law, but for which no other differences were found:

11. The flow of thought in this text is illogical at some points (*original participants agree less than no ZL-participants*)
18. This text is monotonous (*original participants agree less than no ZL-participants*)
28. This text has a messy structure (*original participants agree less than no ZL-participants*)

40. This text has a clear structure (*original participants agree more than no ZL-participants*)

It seems to be that the structure of the text is somewhat obscured in the text without Zipf's law compared to the original text. The control text is in the middle in this respect but does not differ from the other two texts.

Questions on which the original text is rated differently than the control text, but for which no other differences were found:

6. I found this text difficult to finish (*original participants agree less than control participants*)
17. This text is a quick read (*original participants agree more than control participants*)

The control text was considered to be more difficult to get through than the original text. The text without Zipf's law was in the middle but did not significantly differ from the other two texts in this respect.

There was one question for which the text without Zipf's law and the control text were rated differently than the original text, but for which no other differences were found:

2. This text contains many difficult sentences (*control participants agree more than no ZL-participants*).

The original was in the middle but did not significantly differ from the other two texts in this respect.

Questions on which no differences were found:

2. This text contains many difficult sentences.
4. This text contains many ambiguities.
9. The sentences in this text are too long.
10. This text can be considered literature.
13. This text forms a unity.
14. Word choice and formulation seem more important than content.
15. This text is enthralling.
16. This text is picturesquely written.
19. The author could have said the same thing with fewer words.
20. The sentences in this text are too short.
21. This text is unambiguous.
22. The individual paragraphs in this text are in a logical order.

23. This text has a clear story.
25. This text has a pleasant rhythm.
30. This text is boring.
32. The individual paragraphs in this text have little to do with each other.
33. This text is dull.
35. This text is written by an amateur.
37. This text is varied.
38. This text does not contain a clear story.
41. There is not a sentence too much in this text.
42. This text is funny.

Part C | Question 43 – 45: Variance in word choice

For all three questions regarding variance in word choice it was found that participants detected the manipulations that were applied. Both for verbs (Q43), nouns (Q44) and for adverbs and adjectives (Q45), word choice was rated as less varied for the no Zipf's law-text than for the original text. For nouns, word choice of the no Zipf's law-text was also rated as less varied than that of the control text. For *p*- and Z-values, see Table 3.21.

Table 3.21. Mean values, p- and F-values for the questions concerning word variance for which significant differences were found. Df1 = 2 in all cases.

		Mean			**Kruskal**		**Z-values post hoc test**			***p*-value post hoc test**		
Q		**original**	**noZL**	**control**	**X^2**	**p**	**cont-noZL**	**or-noZL**	**or-cont**	**cont-noZL**	**or-noZL**	**or-cont**
43	mean	4,774	3,927	4,444	7,915	0,019	-1,621	-2,717	-0,871	0,158	*0,020*	0,384
	median	5	4	5								
44	mean	5,355	4,093	5,185	23,753	0,000	-3,701	-4,262	-0,334	*0,000*	*0,000*	0,738
	median	5	4	5								
45	mean	5,290	4,182	4,885	13,364	0,001	-2,094	-3,530	-1,107	0,054	*0,001*	0,268
	median	5	4	5								

Part D | Question 46 – 49: Errors

In response to these questions, participants reported many errors that were technically not wrong. Participants preferred a different word order or would prefer to use words in a slightly different way. The errors themselves were therefore not analysed. Counts how often participants reported no, some or many errors are given in Table 3.22. The outcomes of the Kruskal-test are given in Table 3.23.

For Q47 concerning errors in word choice and word usage, it is found that participants reported more errors for the control text than for the original text. The no Zipf's law-text is in between but does not differ from the other two texts. For Q49 concerning interpunction, both modified texts differ from the original text, but the two modified texts do not differ from each other.

Part E | Question 50: Text rating

The original text was on average rated with a 6,8, the two adapted texts were both rated with a 5,7. A one-way analysis shows that this difference is significant (F(2,108) = 6,513, *p* = 0,002). A Tukey's HSD post-hoc test shows that the two adapted texts do not differ, but both were rated worse than the original text to the same degree (original – no Zipf's law: *p* = 0,003; original – control: *p* = 0,011).

Table 3.22. Number and percentage of participants who answered "no", "yes, some" and "yes, many" to the question if there were errors in this text.

			No		**Yes, some**		**Yes, many**	
		N valid	**N**	**%**	**N**	**%**	**N**	**%**
46	No ZL	52	31	59,6	19	36,5	2	3,8
	Original	28	18	64,3	10	35,7	0	0,0
	Control	26	15	57,7	9	34,6	2	7,7
47	No ZL	52	23	44,2	26	50,0	3	5,8
	Original	29	18	62,1	9	31,0	2	6,9
	Control	26	9	34,6	9	34,6	8	30,8
48	No ZL	53	50	94,3	3	5,7	0	0,0
	Original	29	27	93,1	2	6,9	0	0,0
	Control	24	22	91,7	1	4,2	1	4,2
49	No ZL	52	22	42,3	26	50,0	4	7,7
	Original	29	23	79,3	5	1,7	1	3,4
	Control	25	8	32,0	7	28,0	10	40,0

Table 3.23. Mean values, medians, p- and Z-values for the error questions.

		Mean			**Kruskal**		**Z-values post hoc test**			***p*-value post hoc test**		
Q		**original**	**noZL**	**control**	**X^2**	**p**	**cont-noZL**	**or-noZL**	**or-cont**	**cont-noZL**	**or-noZL**	**or-cont**
46	mean	0,357	0,442	0,500	0,506	0,777	-	-	-	-	-	-
	median	0	0	0								
47	mean	0,448	0,615	0,962	6,740	0,034	-1,718	1,233	2,586	0,129	0,218	*0,029*
	median	0	1	1								
48	mean	0,069	0,057	0,125	0,230	0,891	-	-	-	-	-	-
	median	0	0	0								
49	mean	0,241	0,654	1,080	17,168	0,000	-1,949	2,773	4,093	0,051	*0,008*	*0,000*
	median	0	1	1								

Part F | Question 51 – 53: Other

The questions 51 ("Are there other things that strike you about this text which were not addressed in this questionnaire?") and 52 ("Is there anything else that you would like to mention about this text or questionnaire?") were included to see if there were any unforeseen problems with either questionnaire or text. This was not the case.

3.3.3 Discussion

The aim of this study was to find out if it is possible to create a text that does not conform to Zipf's law, what such a text looks like and to see how readers evaluate such a text.

To this aim, I took a text and modified it in two ways: one version of the text was constructed such that it should not conform to Zipf's law, and one version was adapted in other ways while leaving Zipf's law intact. The non-Zipfian text was created by first calculating an artificial word frequency distribution to which the lexical words in the adapted text should comply. This artificial distribution had the same mean frequency as the lexical words in the original text, but word frequencies were made to follow a log-normal distribution on Zipf's β-law instead of a log-linear distribution. Grammatical words were not manipulated to prevent ungrammatical sentences. The control text, meant to test the effect of manipulations of Zipf's law against manipulations in general, was constructed by changing words for their (near) synonyms, changing sentence length and by replacing unstressed pronouns for their stressed counterparts.

The word frequency distribution in the resulting non-Zipfian text was clearly disrupted. For lexical words, Zipf's law was arguably absent: the parameter values ZM-α = 2,53 and especially ZM-β = 204,42 indicate a curve instead of a log-linear distribution, and the fit to especially Z-β is very low, at only Z-R^2 = 0,53. No clear definitions exist of when Zipf's law does not apply, but the abnormality of these values suggest that this conclusion can safely be drawn nonetheless. The lack of disruptions in the word frequencies of grammatical items somewhat obscures the deviations from Zipf's law when the full text is considered, causing parameter values to fall within the normal range. Still, however, the fit to Zipf's β-law is with Z-R^2 = 0,78 very low (usually, values well above 0,9 are found). In addition, much fewer hapax legomena are present. Disruptions of the frequency distribution thus clearly exist, although without a proper definition it can be debated if Zipf's law is truly absent.

These texts were read by three different groups of participants, who were asked to fill in a questionnaire about these texts. The questionnaire was designed to address as many text characteristics as possible, since no expectations were present about the kind of differences readers would experience.

The questions concerning variance in word choice revealed that participants noticed the changes made to the non-Zipfian text. Both for verbs, nouns and for adverbs and adjectives, word choice was rated as less varied for the no Zipf's law-text than for the original text. For nouns, word choice of the no Zipf's law-text was also rated as

less varied than that of the control text. These findings are a direct result of the manipulations that were applied.
The two adapted texts were comparable, judged by the fact that they were both rated with a 5,7 on a scale from 1 to 10, which was significantly worse than the 6,8 given to the original text. The reason for this lower mark, however, differed per text.

Generally speaking, the original text was rated as more natural, and more pleasant and easy to read than the two adapted texts. It seems to be that the structure of the text was somewhat obscured in the text without Zipf's law compared to the original text. The control text is in the middle in this respect but does not differ from the other two texts. The control text was considered to be more difficult to get through than the original text. The text without Zipf's law was in the middle but did not significantly differ from the other two texts. The manipulations of the control text were noticed, as shown by the fact that this text was rated worse on questions concerning strange, aberrant or old-fashioned word use and strangely constructed sentences.

The only question on which the text without Zipf's law stands out is the question concerning how poetic the text was. The text without Zipf's law was rated as less poetic than the other two. It is unclear why this is the case. It might have to do with the many word repetitions. Poetry is known for creative use of words. The text without Zipf's law was very un-creative in this way, because the same words were used more often than usual. Poetry is also often characterized by some form of rhythm. This might be another reason why the non-Zipfian text is rated lower on this aspect. The rhythm that is distorted is that of information density. In normal texts, there is an alternation between high and low entropy words, in other words, words that carry much or little information. Speakers unconsciously aim to keep the amount of information transmitted per unit of time more or less uniform (Frank & Jaeger, 2008; Levy & Jaeger, 2007). Very specific words are more informative and thus higher in entropy than very general words. The manipulation of the text consisted of the exchange of low frequency items for high frequency items. This often resulted in the substitution of specific words for more general ones. It is highly likely that this process has altered the information density of the text. Speakers might unconsciously dislike this. More research is necessary to unveil the exact reasons for this difference.

The error questions were somewhat difficult to interpret. Many participants reported errors that were technically not wrong. A reason for this might be that the questions triggered them to actively search for errors, even though they were not actually present. Differences in the amount of errors reported were found for the questions concerning word choice and usage, and for interpunction. For the question concerning errors in word choice and word usage, it was found that participants reported more errors for the control text than for the original text. The non-Zipfian text is in between but does not differ from the other two texts. This is probably a result of the manipulations applied, namely switching words for their (near) synonyms. This could have resulted in less preferred words in these positions. It is interesting to see that similar manipulations did not have the same effect for the non-Zipfian text, possibly because in that text words were replaced by more general terms. Apparently, this was less often perceived as erroneous. For the question concerning interpunction, both modified texts differ from the original text, but the two modified texts did not differ from each other. This is mainly due to reported missing comma’s. Apparently, I prefer fewer comma’s in my texts than most participants in this study.

Only very few questions elicited different responses for the two adapted texts. This might be because the non-Zipfian text was not disrupted enough. As discussed above, Zipf's law was absent from lexical words, but not from grammatical words and as a result only partly from all words. It might be the case that the manipulations that were applied were not strong enough. It was, however, the best way I could think of to construct such a text that would not result in a syntactically aberrant text. In any way, this study was to my knowledge the first to even attempt removing Zipf’s law from a text. The method applied here is only one possible method, it might be possible to think of a different approach that does disrupt Zipf's law for lexical and grammatical words equally.

3.4 General discussion

Zipf's law has been known for over a century, and yet much is unknown about its exact workings. Zipf's law has hardly been systematically studied in small samples, for different genres, for different mediums or for different (especially very short) text lengths. Without attempting to answer all questions, this study can hopefully help to gain more insight at these points. It is necessary to know more about Zipf's law under these specific conditions, before studying Zipf's law in special cases such as aphasic speech.

For the first part of this chapter I took texts on a continuum from spontaneously spoken language to extensively thought through written language. The categories of texts included were face-to-face conversations, spontaneous commentaries on radio or television, radio and television discussions, sermons and speeches, blogs, news articles and literature. Conversations, commentaries, discussions and sermons/speeches were considered spoken language; blogs, news articles and

literature were considered written language.
For all categories, growth curves were constructed. ZM-α was found to increase according to a logarithmic function when text size increased. The development of the parameter value, in other words, the shape of the distribution, was found to be identical for all categories. Something very different was found for ZM-β. No pattern could be discovered in the shape of the distribution, meaning that the development of ZM-β when text size increased was different for each text and category. This might be due to the fact that ZM-β is always calculated based on only a handful of observations, thus rendering the value prone to fluctuations. The results for Z-β were somewhat mixed. It seemed like the different growth curves were rather similar: Z-β either decreased slightly when text size grew, or it stayed relatively stable. Nevertheless, no pattern could be discovered when distributions were compared statistically.
Based on the growth curves of ZM-α and Z-β, it was determined that the minimal workable text size for Zipf's law lies around 1500 tokens. To be on the safe side, it was advised to analyse at least 2000 tokens when studying Zipf's law, and to always compare texts of equal length.

Following these findings, the values of the parameters of Zipf's law were calculated and compared for fixed text lengths. Not all texts reached 1500 tokens, so values were compared for 1000, 1500 and 2000 tokens to see if any pattern arises.
Generally speaking, it was found that spoken text had the higher ZM-α and thus lower Z-β values than written text. Higher ZM-α and thus lower Z-β values point to less varied word usage: frequent words are used more frequently, and the total number of tokens is lower. This is exactly the difference one would expect to exist between spoken and written texts, which is also reflected by the difference in type/token ratio between the two categories: generally speaking, size-independent TTR's are smaller for spoken texts than for written texts (higher values indicate more different types in relation to the number of tokens). Zipf's law thus correctly reflects this difference in a measurable way.

In the second part of this chapter the behaviour of Zipf's law was studied from a very different perspective. It has never before been attempted to remove Zipf's law from a text, while at the same time keeping the text readable. This question first requires a definition of when Zipf's law does not apply, and the search for an approach to accomplish the removal of Zipf's law.
Deviations from Zipf's law are better detected when Zipf's β-law is considered than when Zipf-Mandelbrots law is considered, due to the absence of a ranking variable (see also Chapter 2, section 2.2.1). The ranking that is part of Zipf-Mandelbrots law forces the distribution to have a declining slope. Deviations can be detected when visually inspecting the distribution, but they do not necessarily show up in the

values of the parameters or the fit of the distribution. This is different for Zipf’s *β*-law. This formulation of the law is based on the size of the frequency classes, so in theory this distribution can take any shape. In practice, it follows a log-linearly declining distribution when Zipf's law applies. This typical distribution was disrupted here to obtain the non-Zipfian text: an artificial word frequency distribution was calculated in which the size of the word frequency classes followed a log-normal distribution instead. The text was then manually adapted to fit this new word frequency distribution. This was done for lexical words only: disrupting the distribution of grammatical words would disrupt syntax, thus making it impossible to know what is being measured. The result of these manipulations was that the value of Z-*β* was about half that of the normal value for texts of this size (Z-*β* = 1,4 while Z-*β* = 2,2 for the original text). The biggest pointer to a disrupted distribution, however, was the fit to Zipf’s *β*-law, Z-R^2 = 0,78, while usually values well above 0,9 are found. The values for Zipf-Mandelbrots law did not reveal the disruptions, although differences in the number of hapax legomena were visible from the plots of Zipf-Mandelbrots law. The calculation of Zipf-Mandelbrots law for lexical words only did show a difference: the value for ZM-*β* was much higher than usual (ZM-*β* = 204,4, while normally values above 10 are rare), and ZM-*α* = 2,5, more than six times as high as that of the control text (ZM-*α* = 0,4). These differences thus disappeared when lexical and grammatical words were combined in the analyses.

What this shows is that Zipf's *β*-law can be disrupted while Zipf-Mandelbrots law still displays rather normal parameter values. Looking at parameter values alone is thus not enough when this formulation of Zipf's law is concerned, and even the fit in terms of R^2 does not always reveal the disruption. The plot of Zipf-Mandelbrots law does show disruptions for the lower frequencies, but their influence on the values of the parameters is small. Zipf’s *β*-law is much better at revealing the disruptions. The fit is reduced to almost 80% of its usual value, and the value of Z-*β* is halved. It is thus advised to use Zipf’s *β*-law instead of Zipf-Mandelbrots law whenever the question arises if Zipf's law applies. This is important information when Zipf's law is studied in aberrant populations, such as speech from people with aphasia. This will be discussed in the next chapter.

People who read the non-Zipfian text did notice that the variation of words was smaller than that in normal texts. They did not particularly like the text but did not dislike it more than when the text was modified in a different way. It seems to be the case that people simply do not like modified texts, which is actually not very surprising. But the underlying reason for this is unclear: did participants dislike the text because it does not conform with every other text they ever read, or because it goes against their innate system? More research is needed to learn about the exact workings behind these judgments. I will discuss this issue further in the General Discussion in Chapter 6.

The combination of Part I and Part II of this chapter provides some additional insight into the values of the parameters of Zipf(-Mandelbrot)'s law and their relation to the typical shape of the distribution. In Chapter 2 (Section 2.2.1) it was discussed that lexical and grammatical words or possibly open and closed class words might be responsible for the typical curvature of Zipf(-Mandelbrots) law. The reason for this would be that grammatical words are more strongly represented in the higher ranks of Zipf's law, while the lower ranks are mostly filled with lexical words (e.g. Ferrer i Cancho & Solé, 2000; Popescu, Altmann & Köhler, 2010). Both classes of words individually have a different slope, which becomes most clear when large corpora are examined. The first part of the distribution has a slope of approximately ZM-$\alpha \approx 1$, the second part has a slope of ZM-$\alpha \approx 2$ (Ferrer i Cancho & Solé, 2000; Ha et al., 2006). The mixture of the two results in the typical curved shape of Zipf's law.

Such a difference between lexical and grammatical words was also found for the original text that was used here. For lexical words, it was found that ZM-α = 0,413, while for grammatical words it was found that ZM-α = 1,611.[17] The slope of the grammatical words is thus much steeper than that of the lexical words. This is not surprising, given that grammatical words are generally from closed classes. There are simply fewer different words available, and thus fewer ranks. The result is a steeper slope.

If the difference between lexical and grammatical words causes the curved shape of Zipf-Mandelbrots law, then it might be expected that these classes of words separately do not display a curvature. This would mean that for both classes separately, ZM-β would be close to zero. I am not aware of any previous work in this domain, but here, for lexical words, this is indeed the case: it is found that ZM-β = -0,519. For grammatical words, however, it was found that ZM-β = 9,676. For grammatical words, the curvature thus still exists. The reason for this might lie

[17] The text that was used here was only 954 tokens long, and split into resp. 455 and 503 tokens for the analysis of lexical and grammatical words separately. This means that the differences as they were found might change when larger samples are studied. However, it is highly unlikely that the effect completely disappears, or even becomes switched. The values reported here are close to the extremes of the range of values for ZM-α for all parts of speech combined that were found in Part I of this chapter, where values between ZM-α = 0,3 and ZM-α = 1,3 are found for the full range of growth curves (for samples from 100-5000 tokens). The value found for grammatical words only is thus even higher than the extreme of all parts of speech combined.

in the already above-mentioned principle of saturation (the process that forces the frequency distribution into a curve when a text grows without the introduction of new types), because grammatical words are usually from closed classes: this means that for large texts, the point can be reached at which no new types are being encountered.

The presence of a clear curvature for grammatical words casts doubt on the claim that it is the combination of lexical and grammatical words that causes the curvature of Zipf's law. There might be two classes of words, but lexical/grammatical does not seem to be the right dichotomy. The same holds for open/closed class words, since saturation occurs especially for closed class words. An option that remains open is the dichotomy proposed by Ferrer i Cancho and Solé (2000), who argue that words can be split into a kernel lexicon formed by a language dependent number of versatile words (ca. 5000-6000 for English) and the rest of the lexicon for specific communication. However, this dichotomy rather reflects the change in slope that is often observed in very large corpora somewhere between rank 10.000 and 100.000. It is unclear if this dichotomy could also explain the curvature for the first few ranks of Zipf-Mandelbrots law. The texts used here are too small to investigate this claim any further.

It is now known that Zipf's law behaves identical for different text sizes, but with different parameter values for different categories. In addition, it has become clear that Zipf's β-law is more sensitive to disruptions to Zipf's law than Zipf-Mandelbrots law, although Zipf-Mandelbrots law can be studied in somewhat smaller text samples. These findings form a solid background for the study of Zipf's law in special populations, such as speech of people with aphasia. This will be the topic of the next chapter.

3.5 Conclusion

Zipf's law should not be studied for texts shorter than 1500 words, and texts of at least 2000 words are preferred. From this number of tokens onwards, parameter values can be compared for texts with equal numbers of tokens. It is difficult to say when a text does not conform to Zipf's law, but the values of the parameters fluctuate strongly for very small text sizes. Zipf's law is thus not a suitable method to directly compare texts that are so short.

The values of the parameters of Zipf's law are different for spoken and written texts: generally speaking, spoken texts display higher values for ZM-α compared to written texts, and lower values of Z-β. This finding reflects a smaller vocabulary in spoken texts than in written texts, even if a spoken text is prepared beforehand.

It is possible to construct a readable text that does not conform to Zipf's law, at least for content words. Readers dislike this as much as an otherwise manipulated text in which words were changed for their (near) synonyms, sentence length was adapted and non-stressed pronouns were replaced by their stressed counterparts. The only real difference between the two modified texts was that the non-Zipfian text was judged as less poetic, possibly due to the less diverse word usage. More research is necessary to discover why exactly people dislike a text without Zipf's law.

Zipf-Mandelbrots law is not very sensitive when it comes to disruptions of Zipf's law, but Zipf's β-law is. It is therefore advised that, when text size allows it, to use Zipf's β-law whenever doubt arises concerning the applicability of Zipf's law.

These findings provide a background to study Zipf's law in special populations, such as people with aphasia.

4 Zipf's law in aphasia

4.1 Introduction

The parameters of Zipf's law are influenced by text characteristics such as length, number of topics, medium and genre (see Chapter 3), but the general shape of the distribution of word frequencies is identical for all texts for which it has been studied. The question addressed in the current chapter is if Zipf's law continues to hold if the language system is damaged as a result of the language disorder aphasia. The results show that it does.

4.1.1 Zipf's law in aphasia

Brain damage such as a stroke, trauma or tumour can result in aphasia, an acquired language disorder. The set of symptoms of the disorder is diverse and slightly different for each patient, dependent on the exact part of the brain and the language faculty that is affected (for more details, see Chapter 2, Section 2.1.3). Patients can be classified in a number of ways. Here, a broad distinction between fluent and non-fluent aphasia is used. My focus is on non-fluent aphasia, which includes word form aphasia and semantic aphasia (Laine & Martin, 2006). Word form anomia is caused by difficulties in accessing the lexeme that specifies the morphological and phonological information of the target word. This kind of impairment is more well-known under the label of agrammatism. Semantic anomia a form of word finding difficulties usually paired with comprehension difficulties and relatively preserved phonological abilities, caused by conceptual- and lemma-level deficits. The level of concreteness/abstractness of the target word can influence the degree of the impairment, as can imageability. Word-class-specific impairments are also a well-known phenomenon. A function word/content word dissociation, for example, is known to exist in many agrammatic aphasics: many of them struggle more with function words than with content words.

From the discussion of potential hypotheses for Zipf's law in Chapter 2, it followed that the hypotheses explaining Zipf's law through preferential attachment were most plausible. These theories place Zipf’s law in the organization of networks of semantic or possibly lexical knowledge. Significant differences in the parameters or fit of Zipf's law for aphasic speakers compared to healthy controls thus suggests that

Terminology on Zipf's law

The terminology used in this chapter is the same as elsewhere in this dissertation, but is repeated here for convenience and in an attempt to limit confusion to a minimum (see also Chapter 2, section 2.2.1).
The term ‘Zipf's law’ is used as an overarching term for two formulas:

Zipf-Mandelbrot’s law:

$$3)\quad f(w) = \frac{C}{(r(w) + \beta)^{\alpha}}$$

where the frequency *f* of word *w* is determined by its rank *r* when all words are ordered from most to least frequent, the parameters α and β, and a text size dependent constant *C*, and

Zipf’s β-law:

$$4)\quad n_f = C \cdot f^{-\beta}$$

where the number of words *n* with frequency *f*, in other words, the size of the frequency class, is determined by parameter β and a text size dependent constant *C*.

Both Zipf-Mandelbrot’s law and Zipf’s β-law are formulations of the same phenomenon, and the unspecific term ‘Zipf's law’ has throughout the literature been used to describe both. To avoid confusion, when either of the two laws is discussed specifically it will be called by its full name.
Another potential point of confusion is the parameter β, because both laws have one. Therefore, the β from Zipf-Mandelbrot’s law will be dubbed ZM-β, while the β from Zipf’s β-law will be dubbed Z-β.

the storage or retrieval of their semantic or lexical knowledge has been altered. On the other hand, a lack of difference means that it is highly unlikely that this knowledge or its retrieval has fundamentally been affected.

Zipf’s law in aphasic speech, or other impaired populations, has only very little been researched. Word frequency distributions in aphasia in a more general sense were studied by Howes and Geschwind (Howes, 1964; Howes & Geschwind, 1964). They did not use the traditional formulation of Zipf’s law, but used a cumulative version concerning the percentage of words that occur with frequencies up to and including each frequency value. Their results indicate that speech from people with aphasia follows the same distribution as that of healthy speech, albeit with a different slope. Unfortunately, the distribution they studied is not very sensitive: disruptions in the higher frequency classes are easily concealed if the lower frequency classes do follow the Zipfian distribution.

That Zipf's law also holds for non-fluent aphasic speakers did follow from our previous work (Van Egmond, Van Ewijk & Avrutin, 2015; see also Van Ewijk, 2013). We interviewed four participants with chronic aphasia and compared their speech to that of four healthy gender and age matched controls from the Corpus of Spoken Dutch (CGN-consortium, 2014). We found no difference in fit to the traditional Zipf's law for both groups: both had a very good fit above $R^2 = 0{,}95$. We did find a difference in slope: for aphasics, a value of $\alpha = 0{,}834$ was found, while for controls $\alpha = 0{,}677$ was found, which reflects the reduced variability in the vocabulary of the aphasic speakers. However, the speech samples were very short, only 352 tokens long. This is well below the minimum of 1500 tokens that was advised in Chapter 3. The effect could thus be different for longer samples.

Other impaired populations were studied by Piotrovskii, Pashkovskii & Piotrovskii (1994) and Piotrowski and Spivak (2007). They looked at Zipf's law in schizophrenic patients and in children with Down syndrome. They found that Zipf's law still applied, but that the power law had a different slope depending on the conditions of the patient. For schizophrenic patients with disconnected speech they found a more gradual slope (smaller α, reflecting a highly varied vocabulary); for schizophrenic patients with a topic of obsession they found a curve instead of a straight line (reflecting the process of saturation or over usage of the available words, see also Chapter 3, Section 3.2.3), and for children with Down syndrome they found a steeper slope (reflecting a less varied vocabulary). Unfortunately, the length of the text samples differed per person. Nevertheless, these findings suggest that differences might be found for aphasic speakers as well.

4.1.2 Normalised Frequency Difference

The calculation of the parameters of Zipf's law through (a form of) curve fitting is called a parametric measure. This method assumes a pre-defined model, in other words, a formula with one or more parameters such as Zipf-Mandelbrots law with parameters ZM-α and ZM-β, which is fitted to the data by calculating the parameter values for which the model best describes the given data set. More parameters could be added to obtain a better fit. Mandelbrot did this to the traditional Zipf's law when he proposed his version of the law: he added the extra parameter ZM-β to account for the high-frequency curvature. Other versions of Zipf's law with more parameters to better describe the frequency distribution have indeed been proposed, but never became as influential as Zipf-Mandelbrots law (e.g. Naranan & Balasubrahmanyan, 1992).

The downside of parametric measures in general is that more parameters might lead to over-fitting, while using a more parsimonious model might lead to under-fitting.

These problems are solved by avoiding curve-fitting altogether. Such a method was recently proposed by Bentz and colleagues (2017). They present an exact measure for the relative difference between two frequency distributions, the *Normalised Frequency Difference* (NFD). This measure can be used to compare two frequency distributions without the need to first calculate parameters. In this way, two frequency distributions can be compared directly instead of indirectly through the parameter values of a fitted model (e.g. Zipf-Mandelbrots law).

Intuitively speaking, the NFD represents the percentage of token frequency differences per overall number of tokens. It is calculated by dividing the sum of frequencies per rank by the sum of the overall token frequencies for both distributions. More exactly, the measure is calculated as follows.

Let $T = \{t_1, t_2, \ldots, t_V\}$ be the set of word types of size V in a corpus, in other words, the vocabulary, and $F = (f_1, f_2, \ldots, f_V)$ be the distribution of values corresponding to the frequencies of occurrences of each word type in the corpus such that $f_1 = freq(t_1)$. The overall number of tokens N in the corpus C is then the sum of all frequencies:

$$1)\quad N^C = \sum_{i=1}^{V} f_i$$

F^A and F^B are two ranked frequency distributions with vocabulary sizes V^A and V^B. The absolute difference in token frequencies for any given rank i can then be calculated as:

$$2)\quad \Delta Freq(A,B,i) = \begin{cases} |f_i^A - f_i^B| & \text{if } i \leq V^A \wedge i \leq V^B \\ f_i^A & \text{if } i \leq V^A \wedge i > V^B \\ f_i^B & \text{otherwise} \end{cases}$$

In words, this means that the absolute frequency difference is calculated by taking the absolute frequency difference at every rank. If no frequencies are available for either F^A or F^B then it is assumed that the absent frequency value equals zero, which equals taking the frequency of the other vector as the absolute difference.

Based on the frequency difference per rank, the NFD is defined as follows:

$$3)\quad NFD(A,B) = \frac{\sum_{i=1}^{\max(V^A,V^B)} \Delta Freq(A,B,i)}{\sum_{i=1}^{V^A} f_i^A + \sum_{i=1}^{V^B} f_i^B}$$

In words: the sum of all absolute frequency differences is divided by the total number of tokens of the two frequency distributions. To simplify things, the denominator can be substituted with Formula (1):

4) $$NFD(A,B) = \frac{\sum_{i=1}^{\max(V^A,V^B)} \Delta Freq(A,B,i)}{N^A + N^B}$$

Values of the NFD can range from 0 to 1, with values closer to 1 indicating bigger frequency differences. It equals 0 if both frequency distributions are identical. It can only equal 1 if the frequency values of one distribution consists of only zero's while the other consists of at least one non-zero element.

The advantage of this method over the more traditional calculations of the parameters of Zipf's law is that no assumptions are necessary about the underlying distribution, thus circumventing curve fitting.

Bentz et al. argue that for most languages (including Dutch) text sizes of around 10-15 K are required to get a good approximation for these calculations. For the current study, this means that texts from different speakers have to be combined to reach these numbers, thus allowing comparisons between sets but not between individual texts. NFD-calculations are presented here in addition to the more traditional calculations of Zipf's law, which can be calculated for individual texts. The goal is to see if this measure adds any insight in addition to the more traditional parametric measures of parameter comparisons.

4.1.3 Age of Acquisition

Words that are acquired early in childhood are processed faster, more easily or more accurately than words that are acquired late (e.g. Juhasz, 2005). As discussed in Chapter 2 (Section 2.1.2), the three most likely explanations for the AoA-effect are the Semantic Locus Hypothesis, the Network Plasticity Hypothesis and the Lexical-Semantic Competition Hypothesis. According to the Semantic Locus Hypothesis, later learnt concepts are built upon earlier ones. This means that it is the order of acquisition that causes the AoA effect. In terms of Levelt's model (Chapter 2, Section 2.1.1), the AoA effect would then arise at the step of conceptual planning. The Network Plasticity Hypothesis assigns the effect to the structure of the lexical network: early acquired words are learnt when the plasticity of the network is largest, allowing them to shape the network. This would mean that the AoA-effect arises at the step of conceptual preparation and all steps that require access to the lexicon. The Lexical-Semantic Competition Hypothesis divides the AoA-effect into a frequency related and a frequency unrelated AoA-effect. The first stems from the same learning process, the second from competition when the correct lemma for a certain concept must be selected. The effect thus arises at the module of lexical selection.

It thus seems to be likely that the AoA effect originates in the conceptual module of the language production system or at the interface between conceptual preparation and lexical selection. This could (at least partly) be due to early words shaping the network structure when its plasticity is largest.

As mentioned above, the hypotheses explaining Zipf's law through preferential attachment were most plausible. These theories place Zipf's law in the organization of networks of semantic or possibly lexical knowledge. Due to the way in which they developed over time, these networks are structured such that they possess a scale-free small-world structure. Scale-free refers to the property that some nodes (which represent concepts or lemmas) are highly connected, while most have only few connections. As a result, some nodes function as hubs: they are very well connected and travelling from one node to another is highly facilitated by these hubs. What follows from this is the small-world structure, the property that the distance between any two nodes in the network is surprisingly small relative to the size of the network as a whole.

The preferential attachment hypotheses reserve a large role for the age at which a word is acquired. Early acquired words shape the form of the network and are more likely to be more deeply embedded in the structure. This would also render them more resilient to damage to the network. If this is true, then it is expected that people with aphasia display more frequent usage of early acquired words as opposed to late acquired words.

I will test this hypothesis by studying AoA in the spontaneous language of people with and without aphasia. AoA-norms are obtained for the lexical words that participants use. AoA in spontaneous speech has to our knowledge not been tested before.

AoA is highly correlated with frequency, but AoA effects are usually found to be more robust than frequency effects (see Juhasz, 2005, for a general review; and Ellis, 2006, for a review focussing on AoA in aphasia). While AoA is thought to function mainly at the level of conceptual knowledge, frequency of encounter in a language in general is thought to operate on multiple levels of speech production, possibly more pronounced at the level of phonological retrieval (see also Chapter 2, Section 2.1.2). The interplay between Zipf's law and general frequency is currently unknown. I therefore also include an analysis of the distribution of general frequency in Dutch in spontaneous speech.

4.1.4 The current chapter

Following the discussion above, the current chapter addresses the following research questions:

1. Are there any differences in the distribution of Zipf's law in speech from people with and without aphasia?
2. Is there a difference between the distributions of AoA and/or frequency in the Dutch language in general in the speech of people with and without aphasia?
3. In addition, does any of the measures above correlate with the severity of aphasia, and do we see any developments when people recover?
4. Does Bentz and colleagues' NFD measure add anything to the picture as it follows from the questions above?

A corpus of long samples of spontaneous speech from people with aphasia is to our knowledge not available. An additional goal of this study was therefore to compose such a corpus.

4.2 Methods

4.2.1 Participants

The participants in this study consisted of two groups: seven non-fluent aphasic participants with word finding difficulties, and seven healthy control participants, matched to the aphasic participants on gender, age and level of education.

Aphasic participants

The test group consisted of seven people with aphasia, who suffered from word finding difficulties. All of them suffered from a unilateral lesion resulting from a cerebrovascular accident in the left hemisphere. All patients had normal or corrected to normal vision and hearing, and did not suffer from neglect, hemianopsia (decreased vision or blindness in half the visual field) or other visual problems. All of them were native speakers of Dutch, and they had no known language or speech impairment before the acquired brain damage. Mean age was 60,7 years (SD 5,7 yrs.). An additional participant (A01) was excluded after the first session, because she suffered from a form of fluent aphasia with hardly any word finding difficulties. Aphasic participant information is provided in Table 4.1.

The aphasic participants were recruited from three Dutch rehabilitation centres: Via Reva in Apeldoorn (1 participant), Sophia Revalidatie in Zoetermeer (1 participant) and Libra Revalidatie & Audiologie in Tilburg (5 participants). All participants were first approached by their local speech therapist if she thought a patient might be a suitable participant. I then contacted them if they agreed to this.

Aphasic participants were diagnosed by their local speech therapist. Most rehabilitation centres (and all rehabilitation centres that participated in the current study) standardly administer the Akense Afasie Test (AAT) to their patients. Aphasic participants in this study were asked for permission to request the results from this test from their therapist for classification purposes, which was granted in all cases. These outcomes are provided in Table 4.2. In addition, they were estimated by their speech therapist to be able to produce a sufficient number of words in spontaneous speech. This automatically entails that the severity of their aphasia was somewhat limited. And although they were referred to me as people with aphasia, four of them were not considered aphasic by the generally used cut-off criterium of having more than 9 errors on the Token Test, part of the AAT (although these four all fall out on other parts of the AAT, thus confirming their language impairment).[18] This means that in some cases, the participants that were included in this study into the group of aphasics actually turned out not to be aphasic according to the stricter norms. This is unfortunate, especially in such a small-scale study. It

Table 4.1. Aphasic participant information

	Gender	**Age at A-1**	**Aetiology**	**Profession/level of education**
A02	m	62	iCVA-left	Intermediate Vocational Education (Installation technician)
A03	f	49	hCVA-left	Executive secretary
A04	m	58	iCVA-left	Intermediate Vocational Education (Nursing)
A05	m	65	iCVA-left	Intermediate Vocational Education (Laying parquet)
A06	m	65	iCVA-left	Bus driver
A07	f	63	iCVA-left	Housekeeping school + partly School of higher general secondary education
A08	f	63	iCVA-left	Housekeeping school

[18] All aphasic participants were referred to me by their local speech therapists as being aphasic. Only later did I discover that, according to the strict definitions, they were not (although all of them had clear language impairments). However, finding participants in the first place was so cumbersome that I did not have the luxury to exclude half of them, or have the time to find others. With the available time and resources, this was the best I could do.

Table 4.2. Outcomes of the Akense Afasietest (AAT) of the aphasic participants. The scores the individual test items are given, and for the Token Test (TT), Repeating, Writing, Naming and Comprehension, the percentiles (P) and stanines (SN, test scores converted to scale from 1 (most severe) to 9 (not disrupted)) are given. Spontaneous speech is rated on a scale from 1 (most severe) to 5 (not disrupted).

	Spontaneous speech						**TT**			**Repeating**			**Writing**			**Naming**			**Comprehension**		
	I	**II**	**II**	**IV**	**V**	**VI**	**Errors**	**P**	**SN**	**Score**	**P**	**SN**	**Score**	**P**	**SN**	**Score**	**P**	**SN**	**Score**	**P**	**SN**
A02	4	3	4	4	3	3	5	93	8	119	62	6	27	28	4	104	87	8	58	16	3
A03	3	5	5	4	3	3	42	30	3	123	67	6	83	88	7	62	38	4	89	59	5
A04	3	4	4	4	4	3	3	97	8	150	100	9	90	100	9	118	100	9	118	99	9
A05	3	4	4	4	3	3	11	84	7	150	100	9	85	93	8	108	92	8	99	79	7
A06	4	5	4	4	4	4	5	93	8	139	85	7	83	88	7	108	92	8	117	99	9
A07	4	4	5	5	5	5	5	93	8	146	96	8	84	90	7	120	100	9	n.a.		
A08	3	4	5	4	4	4	33	50	5	142	91	8	79	81	7	73	46	5	n.a.		

Legend

Meaning of level

Spontaneous speech

I : Communicative behaviour
II : Articulation and prosody
III : Automated language
IV : Semantic structure
V : Phonematic structure
VI : Syntactic structure

Severity scale

Other subtests

1 - 3 : Severe
4 - 5 : Moderate
6 - 7 : Mild
8 - 9 : Minimal/absent

Spontaneous speech

1 : Severe disorder; unintelligible even when listening carefully
2 : Moderate disorder; partly intelligible when listening carefully
3 : Mild disorder; articulatory, phonetic and/or prosodic disruptions are noticed, but overall the utterance is comprehendible
4 : Minimal disorder; articulatory, phonetic and/or prosodic disruptions are only noticed when listening carefully
5 : No disorder

could be that the inclusion of more, and more severe, aphasic participants would have rendered stronger results. On the other hand, while I acknowledge that the severity of the aphasia was limited for the patients included, they were not comparable to the healthy speakers. All aphasic participants in this study had a stroke, and they were all undeniably suffering from language impairments due to this stroke. At the time of inclusion, they were all receiving treatment from a language therapist at a rehabilitation centre.

Participants in this study gave permission to use their speech for quantitative analyses but did not give permission for publication of recognisable utterances. For that reason, no examples of aphasic speech are included here or elsewhere in this study.

The results of the aphasic participants are presented in two ways. First, they are all grouped together and as such compared to the control participants. Second, they are split into mild aphasic participants (AM) and severe aphasic participants (AS; although it should be kept in mind that this division is relative: truly severe aphasic speakers would have been unable to participate in the current study). Participants A2, A4, A6 and A7 are considered mildly aphasic; participants A3, A5 and A8 are considered (more) severely aphasic.

Control participants

The control group consisted of 7 people without any brain damage or language difficulties (based on self-report). Each control participant was one-to-one matched to an aphasic participant on sex, age (+/- 5 years) and years of education (+/- 2 years). Matching on age and education is necessary to control for approximate vocabulary size (before the stroke). Small differences in age or years of education are not expected to be of large influence on vocabulary size, therefore some differences were allowed. They were all native speakers of Dutch, and they all had normal or corrected to normal vision and hearing. Their mean age was 59,6 years (SD 4,4 yrs.). Control participant information is provided in Table 4.3.

Control participants were recruited through my personal network. None of them was related to me or in other ways well known.

All participants received five euros per testing session. All participants were explicitly informed (both in writing and orally) that they could quit the project at any time. None of the participants made use of this option.

Table 4.3. Control participant information

	Gender	Age	Profession/level of education
C02	m	61	Executor- and contractors education
C03	f	52	Intermediate Vocational Education (lab Leyemburg)
C04	m	55	Intermediate Technical Education
C05	m	61	Intermediate Vocational Education met extra certificates
C06	m	62	Lower Technical Education
C07	f	62	Lower Vocational Education (secretary)
C08	f	64	Housekeeping school

4.2.2 Study Design

Testing intervals

All aphasic participants were tested during their recovery, at 2, 5 and 8 months post onset (+/- 2 weeks). These intervals all fall within the post-acute stage of recovery. At this stage, any problems with basic functioning due to the trauma are (being) treated. The nature and severity of their aphasia has been determined, but the patient is still recovering to a large degree (Springer, 2008).

For healthy controls, no significant differences are expected between testing sessions, because language capacities of healthy speakers do not generally change over the time of six months.[19] For efficiency reasons, these participants were therefore tested only once.

Testing procedures and tasks

Participants were tested in a quiet room, at the location most convenient to them. In practice, this was in most cases at their home, and in a few cases at their rehabilitation centre.

All subjects were subjected to the Ruff Figural Fluency task (RFFT), followed by the experimental tasks. The Ruff Figural Fluency Test is included to test for non-verbal cognitive differences between groups and was therefore administered to all participants. This test was administered to the aphasic participants at every testing

[19] Nevertheless, it would have been better to test control participants three times too to control for normal variation over testing sessions. Unfortunately, with the available time and resources, I was unable to handle the additional amount of data.

moment. Testing time is 5 minutes (60 seconds for every sub-part). All participants performed within normal limits on this task.

The experimental tasks were all aimed at eliciting spontaneous speech. These tasks were a picture description task, a film-retelling task and an unstructured interview (e.g. stroke narratives). Both the film-retelling task and the picture description task were used to elicit spontaneous speech while at the same time controlling for the topic of the conversation. Films used for this purpose should not contain any speech in order to allow participants to choose their own words. One film that has frequently been used for this purpose is a fragment from the Charlie Chaplin film Modern Times (see Keijzer, 2007 for an overview), which was therefore also the film used in the current study. The method was identical to the method applied by Keijzer (2007). Participants watched a ten-minute excerpt from the film Modern Times. Participants were informed beforehand that afterwards they would be asked to retell the story in as much detail as possible. I did not leave the room while the participant watched the film, but I did not actively engage in watching it with the subject. Participants were free in their method of retelling but were reminded of fragments that were missing from their narratives and prompted to relate the omitted parts.

The pictures used in the picture description task had to be complex enough to elicit a sufficient amount of speech. For this purpose, five pictures were selected that were originally from the Wasgij jigsaw puzzle range (see Appendix D). The pictures show complex situations with a number of objects and characters, but all have one event in the focus of the picture. An informal pre-test confirmed that they were suitable for the current purpose.

In addition to these directed speech elicitation tasks, participants were invited to talk freely in an unstructured interview. For people with aphasia, the first question in the first testing session was usually how they obtained their aphasia. In the second session, I usually asked about an unrelated event of their lives, such as their family or how they met their partner. These were also the questions I usually asked the control participants. In the third session, I usually asked how they were doing now, after therapy has ended and the aphasia has become a part of their lives. I departed from this schedule if we naturally ended up talking about other subjects, because content of the conversation was not relevant for the study.

This interview was kept as close to a natural conversation as possible. This method was chosen because it is the most natural way to produce spontaneous speech. Due to the nature of the task, the topic of conversation differed between participants.

In total, the time required for both test battery and experimental tasks was approximately 1 hour per testing moment.

Data handling

The speech output of the patients on the different tasks was recorded on audio and on video and orthographically transcribed in CHAT-format (MacWinney, 2000). For each participant and each testing moment, there were three conversations: the unstructured interview, the picture description task and the film-retelling task. The output on all three conversations was combined to obtain samples of sufficient length for further analysis. The usage of quantitative methods justifies this, as these methods generalize over content. For all words, the lemma was noted down.

Values for estimated AoA and for SUBTLEX word frequency were obtained for all words for which they were available. First, it was attempted to retrieve these values for the lexeme as used in the text. If this was not available, it was attempted to retrieve the values for the corresponding lemma. These values were obtained through the Dutch Lexicon Project 2 (DLP2, Brysbaert, Stevens, Mandera & Keuleers, 2016). The DLP2 is a large database of lexical decision data, containing data for 30.000 Dutch lemmas. Next to the lexical decision data, these data include word frequency measures, neighbourhood measures, morphological information and syntactical (part of speech) information for the words used. For the current purposes, only age of acquisition (AoA) and SUBTLEX-frequency values were extracted. SUBTLEX is a large database containing 44 million words from film and television subtitles (Keuleers, Brysbaert & New, 2010). It represents spoken word frequencies better than written corpora do. AoA ratings were obtained by asking large groups of university students at what age they thought they had acquired a word. This method has been validated and outcomes are found to have high correlations with more objective measures of AoA (De Moor, Ghyselinck & Brysbaert, 2000; Gilhooly & Gilhooly, 1980).

4.2.3 Analyses and Statistics

Type/token-ratio's

Type/token-ratio's (TTR) are calculated as a way to compare texts independent from Zipf's law. TTR expresses the number of different words relative to the total number of words. Any differences in TTR between transcripts are expected to be reflected by differences in Zipf's law, since both are measures of word frequencies. A size-independent version of TTR was calculated, based on the first 300 words after the first 50 words.

Zipf's law

The formula that is used for the Zipf's law calculations is Zipf-Mandelbrot’s law, provided as Formula 1 in the box at the beginning of this chapter. Both ZM-α and ZM-β are calculated. In addition, Zipf’s β-law is fitted to the data, Formula 2 in the aforementioned box. The values of the parameters are calculated through maximum likelihood estimation (Murphy, 2015), after making a first approximation of the parameters through linear regression (Izsák, 2006).

Two kinds of comparisons are made: within groups comparisons to compare the results within each group on the three different testing moments, and between groups comparisons to compare the results of healthy and aphasic participants for each of the three testing moments. In general, it is expected that any differences between aphasic and healthy participants will decrease during recovery of the aphasic participants. Within groups results will be tested for significance using repeated measures ANOVA’s, or in case of non-normally distributed data using Kruskal-Wallis tests. These tests will be followed by the appropriate post-hoc tests if any differences are found. Between groups results are tested for significance using Student’s t-tests, or in case of non-normally distributed data using Wilcoxon rank sum tests.

Age of Acquisition and SUBTLEX frequency

AoA and SUBTLEX frequencies were analysed for all words for which ratings were available. Only values for the first 2000 tokens were used if the transcripts were longer (this does not mean that 2000 tokens were used, not for all words were AoA and SUBTLEX ratings available). AoA and SUBTLEX ratings were available for as many as 96,4% of all tokens. Number of types and tokens for which values were available are given in Table 4.4 (tokens) and Table 4.5 (types).

AoA and SUBTLEX ratings are examined in a number of different ways: through density plots, by ranking of values from early to late AoA and through comparisons of means.

All calculations and statistical analyses are performed in R.

Table 4.4. Number and percentage of tokens for which AoA and SUBTLEX ratings were available. If possible, ratings for the lexeme (word as used in the text) were used. If these were not available then, where possible, lemma ratings were used.

		N all tokens	Tokens for which values are available					
			Lemma		Lexeme		Total	
			N	%	N	%	N	%
Session 1	All	12.628	1678	13,3	10.543	83,5	12.221	96,8
	AM	7721	1043	13,5	6416	83,1	7459	96,6
	AS	4907	635	12,9	4127	84,1	4762	97,0
Session 2	All	11.966	1663	13,9	9823	82,1	11.486	96,0
	AM	7495	1107	14,8	6051	80,7	7158	95,5
	AS	4471	556	12,4	3772	84,4	4328	96,8
Session 3	All	13.978	1941	13,9	11.530	82,5	13.471	96,4
	AM	8000	1100	13,8	6563	82,0	7663	95,8
	AS	5978	841	14,1	4967	83,1	5808	97,2
Controls	All	14.000	2194	15,7	11.314	80,8	13.508	96,5
	CAM	8000	1222	15,3	6497	81,2	7719	96,5
	CAS	6000	972	16,2	4817	80,3	5789	96,5
All		52.572	7476	14,2	43.210	82,2	50.686	96,4

Table 4.5. Number and percentage of types for which AoA and SUBTLEX ratings were available. If possible, ratings for the lexeme (word as used in the text) were used. If these were not available then, where possible, lemma ratings were used.

		N all types	Types for which values are available					
			Lemma		Lexeme		Total	
			N	%	N	%	N	%
Session 1	All	1689	419	24,8	1108	65,6	1525	90,3
	AM	1283	322	25,1	903	70,4	1224	95,4
	AS	818	207	25,3	564	68,9	771	94,3
Session 2	All	1730	435	25,1	1113	64,3	1545	89,3
	AM	1297	335	25,8	899	69,3	1234	95,1
	AS	766	198	25,8	564	73,6	762	99,5
Session 3	All	1842	483	26,2	1159	62,9	1639	89,0
	AM	1411	356	25,2	907	64,3	1263	89,5
	AS	996	242	24,3	666	66,9	908	91,2
Controls	All	1978	555	28,1	1218	61,6	1770	89,5
	CAM	1333	387	29,0	862	64,7	1249	93,7
	CAS	1169	308	26,3	765	65,4	1073	91,8
All		4067	1096	26,9	2371	58,3	3467	85,2

4.3 Results

4.3.1 Text descriptives

Text descriptives per transcript are given in Table 4.6.

The length of the transcripts differed substantially between sessions and participants. The shortest transcript has a length of 1278 tokens (A08, second session), the longest one is 5195 tokens long (C03). On average, transcripts from aphasic speakers contained 2156 tokens at testing moment 1, 1955 tokens at testing moment 2 and 2826 tokens at testing moment 3, while control transcripts contained on average 3768 tokens. An analysis of variance shows that these values are significantly different from each other (F(3, 24) = 7,12, *p* = 0,001). A Tukey's HSD post hoc test shows that the control transcripts are significantly longer than the aphasic transcripts at testing moment 1 (*p* = 0,005) and testing moment 2 (*p* = 0,002), but not at testing moment 3. On average, the transcripts from the mild aphasic participants were longer than the ones from the more severe aphasic speakers.

A similar pattern is found for diversity in terms of number of produced types. On average, aphasic participants produced 499 different types at testing moment 1, 488 different types at testing moment 2 and 601 different types at testing moment 3, while control participants produced 782 different types. These differences are significant (F(3,24) = 7,77, *p* < 0,001). A Tukey's HSD post hoc test reveals that aphasic participants produced significantly fewer different types at testing moment 1 (*p* = 0,002) and testing moment 2 (*p* = 0,001), but not at testing moment 3. The difference is larger for the more severely aphasic participants. However, this difference might be due to a difference in number of tokens. Size-independent TTR-values were computed to test for this possibility.

The mean size-independent TTR-values for the transcripts in the different sets are TTR = 0,38 for aphasics at testing moment 1, TTR = 0,42 for aphasics at testing moment 2 and TTR = 0,42 for aphasics at testing moment 3; and TTR = 0,39 for

Table 4.6. Tokens, types and size-independent TTR for all transcripts. TTR is based on the first 300 tokens after the first 50 tokens.

	Tokens			**Types**			**TTR***		
Test moment	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
A02	3322	3487	3557	670	686	692	0,340	0,373	0,420
A03	1981	1779	3077	387	428	567	0,393	0,407	0,390
A04	2193	1810	2532	608	595	727	0,400	0,487	0,483
A05	1561	1421	1979	391	361	459	0,337	0,343	0,360
A06	1714	1725	2296	415	447	525	0,410	0,463	0,407
A07	2958	2187	3996	663	561	737	0,427	0,473	0,463
A08	1362	1278	2342	357	337	503	0,377	0,390	0,423
C02	4865			935			0,397		
C03	5195			923			0,423		
C04	3847			818			0,360		
C05	3866			844			0,367		
C06	3325			630			0,350		
C07	2720			711			0,453		
C08	2559			616			0,390		
Mean AM	2547	2302	3095	589	572	670	0,39	0,45	0,44
Mean AS	1635	1493	2466	378	375	510	0,37	0,38	0,39
Mean A (all)	2156	1955	2826	499	488	601	0,38	0,42	0,42
Mean C	3768			782			0,39		

controls. A one-way analysis of variance shows that these values are not significantly different from each other (F(3,24) = 1,41, p = 0,26). The difference in number of types that was found above was thus indeed a result of the larger number of tokens that was produced.

Unfortunately, three of the 28 text samples did not reach the minimally required 1500 words. All analyses hereafter will therefore be applied to three text sizes:

- Shortest number of tokens reached by all samples, which is 1278
- 1500 tokens by those samples that reach this (25 out of 28)
- 2000 tokens by those samples that reach this (18 out of 28)

This will allow us to see if any patterns arise for the longer samples.

4.3.2 Zipf's law

The mean values per testing session are given in Table 4.7. Values per transcript are provided in Appendix E. The distribution of Zipf-Mandelbrot's law is plotted in Figure 4.1, the spread of ZM-α, ZM-β and ZM-R^2 are plotted in Figure 4.2 and 4.3, the distribution of Zipf's β-law is plotted in Figure 4.4, and the spread of Zipf's β-law is plotted in Figure 4.5 and 4.6.

Individual transcripts of aphasic participants are referred to as follows: [participant]:[session] (text size in tokens).

Table 4.7. Mean parameter values for Zipf's law for mild and severe aphasics separately, for all aphasic participants and for control participants, for 1278, 1500 and 2000 tokens.

		ZM-*α*			**ZM-*β***			**ZM-R2**			**Z-*β***			**Z-R2**		
	Sess.	**1278**	**1500**	**2000**	**1278**	**1500**	**2000**	**1278**	**1500**	**2000**	**1278**	**1500**	**2000**	**1278**	**1500**	**2000**
C	1	1,030	1,046	1,109	2,318	2,594	3,727	0,981	0,981	0,985	1,826	1,812	1,807	0,993	0,994	0,993
All A	1	1,041	1,069	1,089	2,482	2,889	3,640	0,977	0,981	0,987	1,799	1,782	1,843	0,992	0,991	0,991
All A	2	1,005	1,029	1,056	2,182	2,721	2,635	0,975	0,982	0,980	1,818	1,843	1,814	0,992	0,993	0,993
All A	3	0,999	1,030	1,081	1,948	2,388	3,118	0,985	0,983	0,981	1,830	1,811	1,765	0,992	0,992	0,991
AS	1	1,101			2,935			0,982			1,708			0,994		
AS	2	1,027			2,297			0,968			1,763			0,990		
AS	3	1,042	1,077	1,130	2,579	2,998	4,016	0,981	0,979	0,974	1,738	1,723	1,670	0,994	0,994	0,994
AM	1	0,996	1,027	1,089	2,038	2,542	3,640	0,976	0,978	0,987	1,838	1,823	1,843	0,992	0,990	0,991
AM	2	0,989	1,005	1,055	2,095	2,284	2,635	0,981	0,983	0,980	1,859	1,866	1,814	0,993	0,993	0,993
AM	3	0,966	0,995	1,059	1,475	1,930	2,669	0,988	0,985	0,984	1,899	1,876	1,813	0,991	0,990	0,990

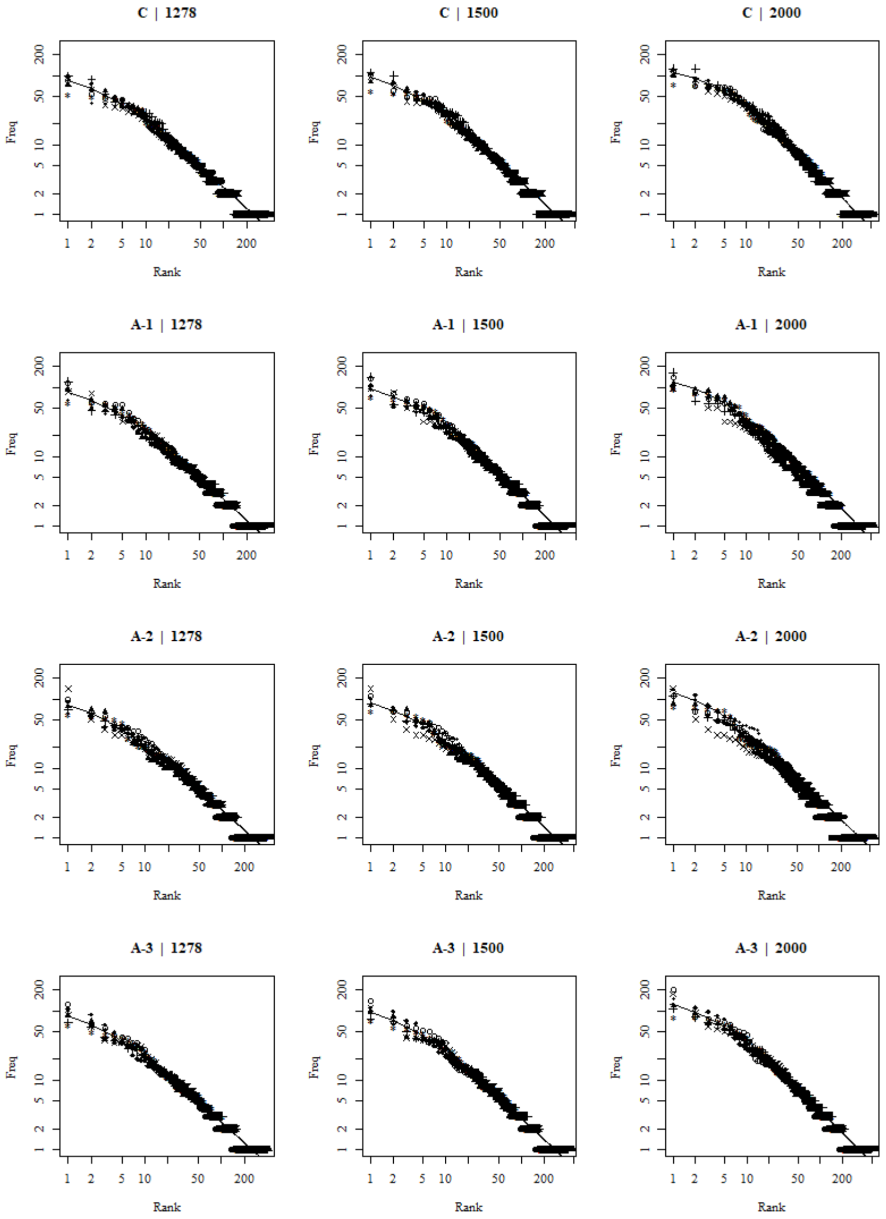


Figure 4.1. Distribution of Zipf-Mandelbrot's law in the transcripts of controls (top row), aphasics session 1 (second row), aphasics session 2 (third row) and aphasics session 3 (bottom row) for 1278 tokens (first column), 1500 tokens (second column) and 2000 tokens (third column). Different symbols indicate different individual texts.

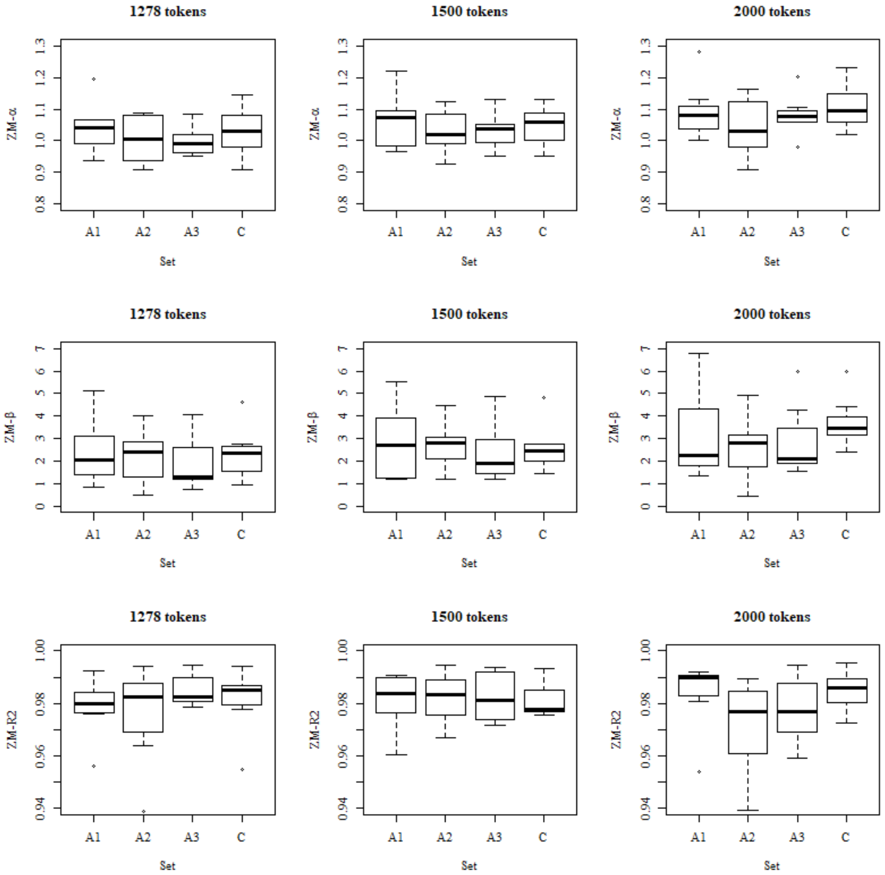


Figure 4.2. Boxplot of ZM-*α* (top row), ZM-*β* (middle row) and ZM-R^2 (bottom row) values for 1278 tokens (first column), 1500 tokens (second column) and 2000 tokens (third column).

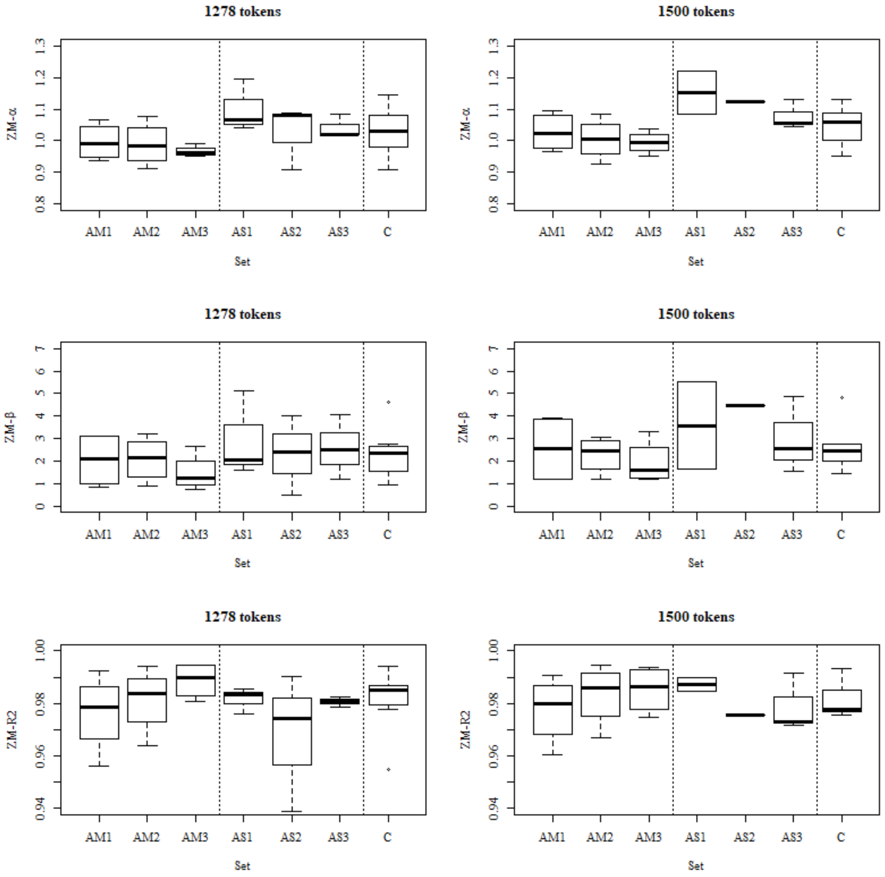


Figure 4.3. Spread of ZM-α (top row), ZM-β (middle row) and ZM-R^2 (bottom row) values for 1278 tokens (first column) and 1500 tokens (second column, divided by mild aphasic participants (AM), severe aphasic participants (AS) and controls (C) (samples from the severe aphasic speakers did not reach 2000 tokens).

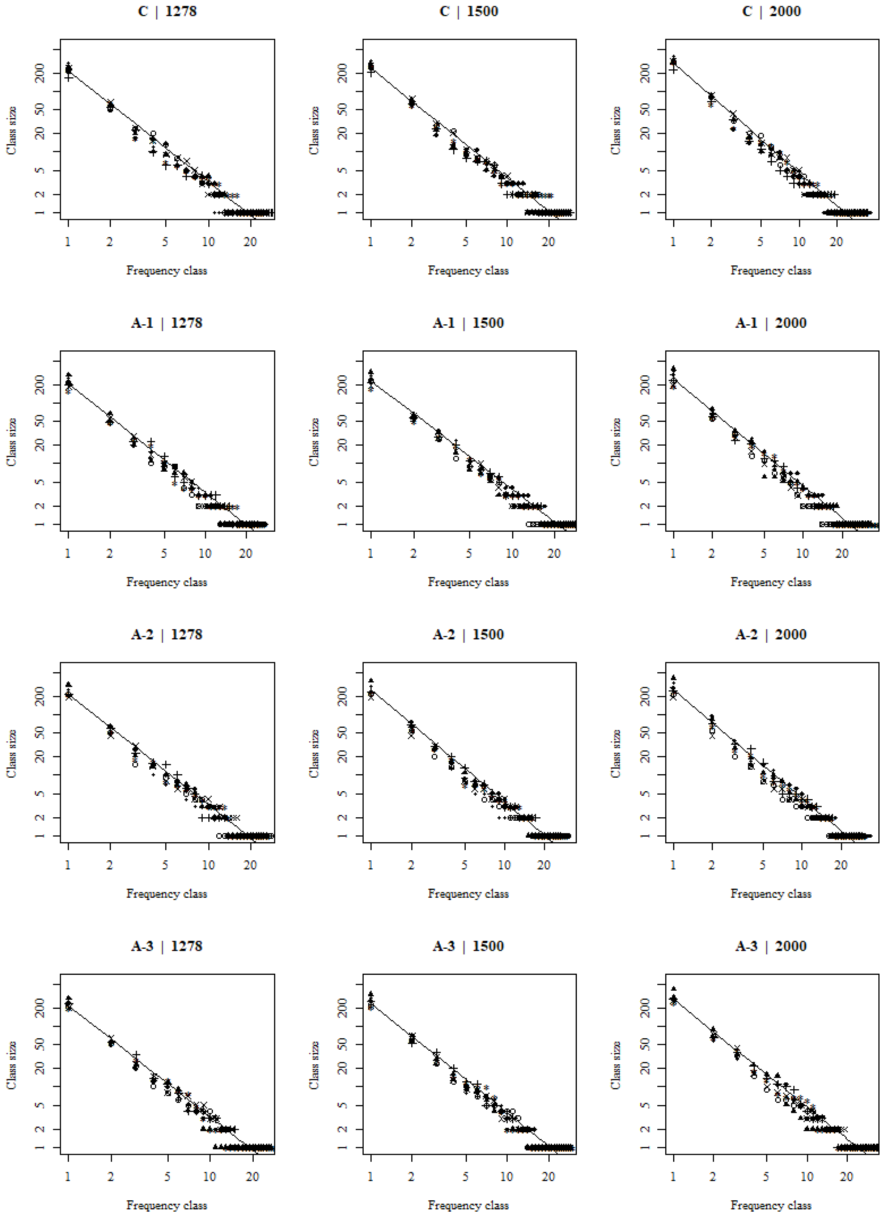


Figure 4.4. Distribution of Zipf's *β*-law in the transcripts of controls (top row), aphasics session 1 (second row), aphasics session 2 (third row) and aphasics session 3 (bottom row) for 1278 tokens (first column), 1500 tokens (second column) and 2000 tokens (third column). Different symbols indicate different individual texts.

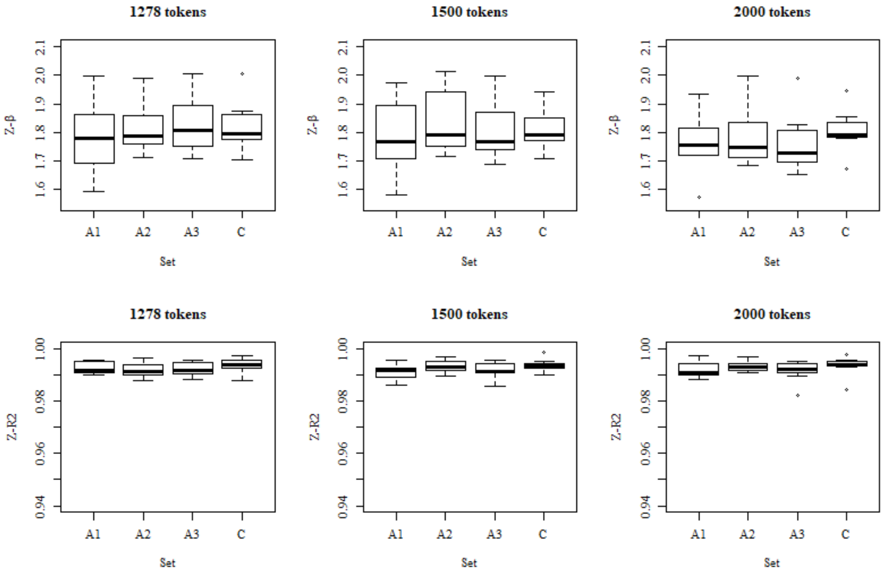


Figure 4.5. Spread of Z-*β* (top row) and Z-R^2 (bottom row) values for 1278 tokens (first column), 1500 tokens (second column) and 2000 tokens (third column).

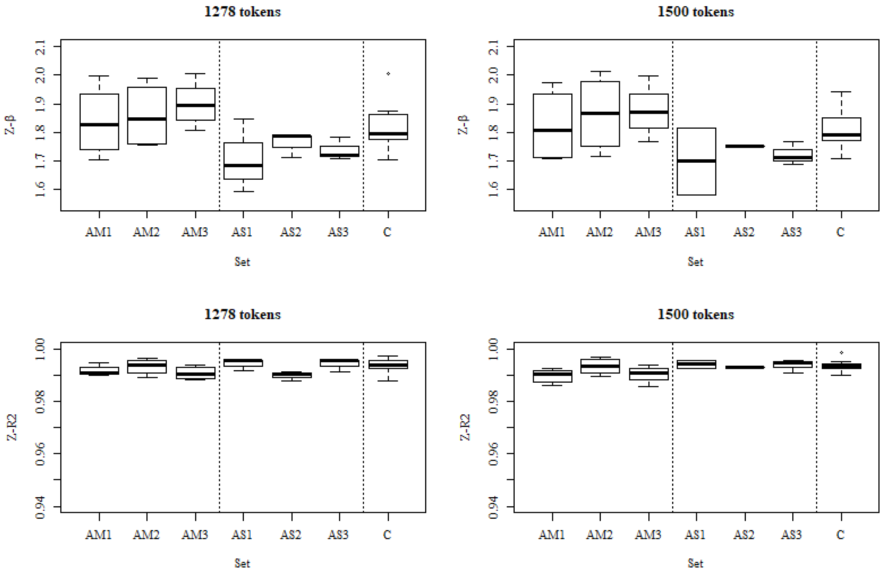


Figure 4.6. Spread of Z-*β* (top row) and Z-R^2 (bottom row) values for 1278 tokens (first column) and 1500 tokens (second column, divided by mild aphasic participants (AM), severe aphasic participants (AS) and controls (C) (samples from the severe aphasic speakers did not reach 2000 tokens).

Zipf-Mandelbrot's law

For ZM-α, the smallest value that is found in any session is ZM-α = 0,910 for A08:2 (1278), the largest value that is found is ZM-α = 1,232 for C06 (2000). For all speakers, ZM-α slightly grows when larger samples are taken into account. This is in line with the findings in Chapter 3, where it was also found that ZM-α grows when text size was increased. On average, for session 1 it is found that ZM-α grows from ZM-α = 1,041 to ZM-α = 1,089 for 1278 to 2000 words; for session 2 it grows from ZM-α = 1,005 to ZM-α = 1,056 and for session 3 it grows from ZM-α = 0,999 to ZM-α = 1,081. For controls, it is found that ZM-α grows from ZM-α = 1,030 to ZM-α = 1,109 for 1278 to 2000 words.

None of the differences between groups for ZM-α are significant, no matter the text size under consideration. Statistically, people with aphasia display a ZM-α indistinguishable from that of people without aphasia.
The difference seems to be somewhat bigger when taking severity into account, especially for session 1. Severely aphasic speakers now seem to display a higher ZM-α than controls or people with mild forms of aphasia. Due to the small sample sizes, this difference could not statistically be compared.

Values found for ZM-β are more diverse, in line with the findings in Chapter 3. ZM-β usually but not always grows when larger samples are taken into account. The smallest value that is found is ZM-β = 0,467 for A08:2 (1278), the largest is ZM-β = 5,955 for both A03:3 (2000) and C03 (2000). On average, it is found that for session 1, ZM-β grows from ZM-β = 2,482 to ZM-β = 3,640 for 1278 to 2000 words, for session 2 it grows from ZM-β = 2,182 to ZM-β = 2,635 and for session 3 it grows from ZM-β = 1,948 to ZM-β = 3,118. For controls, it grows from ZM-β = 2,318 to ZM-β = 3,727. Again, none of the differences are significant: people with aphasia are indistinguishable from people without aphasia in terms of ZM-β. Severity does not seem to make a difference.

The last measure to examine is the fit of Zipf-Mandelbrot's law. The value of ZM-R^2 is high in all cases. The smallest value that was found is ZM-R^2 = 0,955 for C07 (1278), the highest value that was found is ZM-R^2 = 0,995 for A06:2 (1500), A06:3 (2000), C02 (2000) and C07 (1278). It is not necessarily the case that longer transcripts exhibit a better fit. For session 1, average ZM-R^2 ranges from ZM-R^2 = 0,977 (1279) to ZM-R^2 = 0,987 (2000). For session 2, it ranges from ZM-R^2 = 0,975 to ZM-R^2 = 0,982. For session 3, it ranges from ZM-R^2 = 0,981 (2000) to ZM-R^2 = 0,985 (1278). For controls, it ranges from ZM-R^2 = 0,981 (1278 and 1500) to ZM-R^2 = 0,985 (2000).

None of the differences in fit are significant. Both aphasic speech (irrespective of severity) and healthy speech follow Zipf-Mandelbrot's law equally well.

Zipf's *β*-law

Z-*β* is found to vary from Z-*β* = 1,704 (A06:1, 1278 and C06 (1278)) to Z-*β* = 2,015 (A04:2 (1500)). Only little variation is found between samples of different sizes. In Chapter 3, too, it was found that Z-*β* stayed relatively stable when text size increased. On average, Z-*β* ranges from Z-*β* = 1,782 (1278) to Z-*β* = 1,843 (2000) for session 1, from Z-*β* = 1,814 (2000) to Z-*β* = 1,818 (1278) for session 2, from Z-*β* = 1,765 (2000) to Z-*β* = 1,830 (1278) for session 3 and from Z-*β* = 1,807 (2000) to Z-*β* = 1,812 (1500) for controls. None of the differences between groups are significant, nor do the aphasic texts differ between sessions. It does seem to be the case that the severely aphasic speakers have somewhat lower values for Z-*β* than controls or mildly aphasic speakers, which mirrors the findings for ZM-*α*.

The fit of Zipf's *β*-law is high in all cases. The lowest value that is found is Z-R^2 = 0,982 (A07:3 (2000)), the highest value is Z-R^2 = 0,998 (C08, 1500 and 2000). None of the differences in fit are significant, speech from both groups (irrespective of severity) conforms to Zipf's *β*-law equally well at all points in time.

To summarize, aphasic speech samples are indistinguishable from healthy speech samples in terms of Zipf's law, irrespective of the time post onset and the variant of Zipf's law under consideration. People with more severe forms of aphasia seem to have a somewhat steeper slope of Zipf-Mandelbrots law, and thus a somewhat shallower slope of Zipf's *β*-law. This difference does not in any way affect the shape (i.e. fit) of the distributions.

Normalised Frequency Difference

The current transcripts are too short to compare them individually using the Normalised Frequency Difference (NFD) developed by Bentz et al. (2017). The transcripts were therefore combined per set to reach a satisfactory number of tokens for these calculations. All NFD values are given in Table 4.8.

To have an idea of the meaningfulness of the values that were computed, I calculated the NFD for the difference between the combined first and combined second half of the healthy speakers' transcripts. These samples were generally the longest, thus allowing for this comparison. All samples were cut in half, ignoring the middle token in case of an uneven number of tokens. All first halves and all second halves were then combined for NFD-calculations.

Table 4.8. Normalised Frequency Difference values for the combined transcripts for all set-wise comparisons

	Length	A-1	A-2	A-3	C-1
A-1	12618	/	0,036	0,080	0,070
A-2	12013		/	0,106	0,093
A-3	13979			/	0,032
C-1	14000				/

Each half contained 13.189 tokens. The NFD for these samples was 0,048. This means that when all conditions are exactly the same except for the topic of conversation (and possibly fatigue of the speakers), we find a difference of 4,8% between the frequency distributions. The differences found between A-1 and A-2 and between C-1 and A-3 are even smaller than this. The other differences are larger. The largest difference that is found is between A-2 and A-3, which is 0,106 or 10,6%. This difference is larger than any difference between healthy and aphasic speakers. These findings confirm the findings above: it is not the case that the difference between healthy and aphasic speakers decreases during recovery, nor is it the case that the difference between A-1 and A-3 is larger than the difference between A-2 and A-1 or A-3, which would suggest a development during recovery.

4.3.3 AoA and Frequency

The distributions of AoA and of SUBTLEX frequency per set of transcripts are given in Figure 4.7 and Figure 4.8. Each variable is visualised in three different ways: histograms are given that show the frequency of occurrence of values in the transcripts; values are ranked from early to late AoA and from high to low SUBTLEX frequency and displayed using ranks on double logarithmic scales (in analogy to the plots of Zipf's law); and boxplots are given using logarithmic y-axes.

The graphs display the distribution of the combined transcripts up to 2000 tokens per text in each set. The number of words included are given in Table 4.9. The number of types and tokens per participant were already given in Table 4.4 and 4.5 in Section 4.2.3.

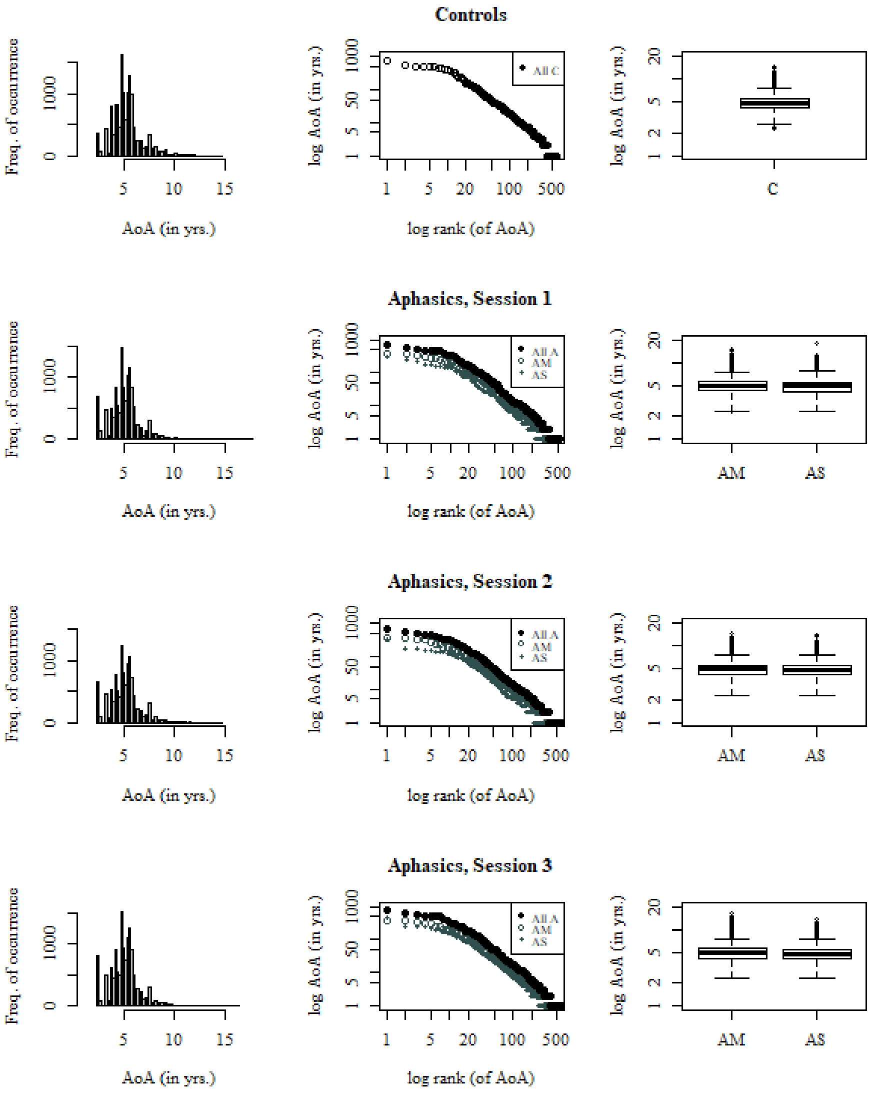


Figure 4.7. Age of acquisition visualisations of the combined transcripts per set. The first column provides histograms of the frequency of occurrence of AoA values; the second column shows log AoA when AoA values are ranked from early to late on double logarithmic scales; the third column shows boxplots of AoA values using logarithmic y-axes.

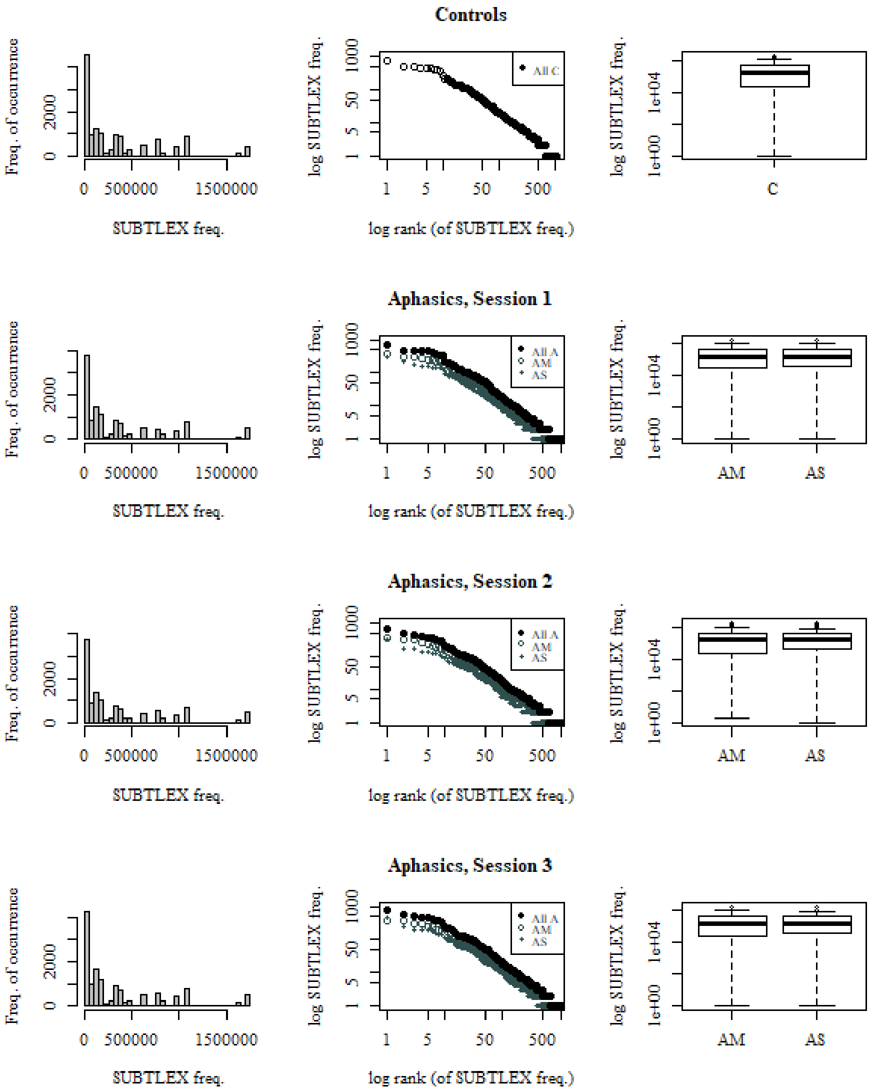


Figure 4.8. SUBTLEX frequency visualisations of the combined transcripts per set. The first column provides histograms of the frequency of occurrence of SUBTLEX frequency values; the second column shows log SUBTLEX frequency when SUBTLEX frequency values are ranked from early to late on double logarithmic scales; the third column shows boxplots of SUBTLEX frequency values using logarithmic y-axes.

Table 4.9. Types and tokens of the combined transcripts per set for which AoA and SUBTLEX frequency values were available. Of each transcript, only the first 2000 tokens were included if the transcripts were longer, or those tokens that were available if shorter.

		Session 1		**Session 2**		**Session 3**	
	Controls	**All aphasics**		**All aphasics**		**All aphasics**	
Types	1.770	1.524		1.545		1.639	
Tokens	13.508	12.221		11.486		13.471	
		AM	**AS**	**AM**	**AS**	**AM**	**AS**
Types		1224	771	1235	762	1263	908
Tokens		7459	4762	7158	4328	7663	5808
		(96,6%)	(97,0%)	(95,5%)	(96,8%)	(95,8%)	(97,2%)

AoA

Summary values for AoA (in years) of all words for which ratings were available per set of combined transcripts (up to 2000 tokens per transcript or as long as available) are given in Table 4.10. Looking at the left column of plots in Figure 4.7, it becomes clear that AoA values are not normally distributed. In all sets, words with an early AoA are more frequent than words with a later AoA. The differences between healthy and aphasic speech are small but significant: people with aphasia tend to use slightly more words with a somewhat earlier AoA (Wilcoxon rank sum tests, A1: V = 34.850.000, A2: V = 34.769.000; A3: V = 34.859.000; all: p < 0,001). This difference seems to be slightly larger for the more severely aphasic speakers. The aphasic samples do not differ per testing moment.

SUBTLEX frequency

Summary values for SUBTLEX frequency of all words for which ratings were available per set of combined transcripts (up to 2000 tokens per transcript or as long as available) are given in Table 4.11. Both groups of speakers display more frequent usage of words with a high SUBTLEX frequency. The distribution of SUBTLEX

Table 4.10. Summary values for AoA (in years) per combined set of transcripts

Group	mean	sd	min	Q1	median	Q3	max
C1	5,21	1,47	2,25	4,34	4,98	5,65	14,60
A1	5,02	1,44	2,25	4,21	4,94	5,61	17,63
A2	5,03	1,46	2,25	4,17	4,94	5,61	14,60
A3	5,02	1,46	2,25	4,17	4,94	5,61	16,25
AM1	5,08	1,47	2,25	4,28	4,96	5,61	14,60
AM2	5,12	1,51	2,25	4,22	5,00	5,67	14,60
AM3	5,11	1,51	2,25	4,22	5,00	5,61	16,25
AS1	4,92	1,38	2,25	4,17	4,89	5,59	17,63
AS2	4,88	1,36	2,25	4,17	4,89	5,52	14,08
AS3	4,91	1,38	2,25	4,17	4,84	5,56	13,89

Table 4.11. Summary values for SUBTLEX frequency per combined set of transcripts

Group	mean	sd	min	Q1	median	Q3	max
C1	340.904,5	429.478,3	0	20.798	150.152	455.234	1.744.062
A1	347.914,6	439.890,0	0	36.366	150.152	443.891	1.744.062
A2	337.205,4	440.869,0	0	30.243	149.900	407.143	1.744.062
A3	341.368,5	438.721,2	0	32.519	149.900	407.143	1.744.062
AM1	348.872,2	446.064,0	0	32.519	149.900	407.143	1.744.062
AM2	338.701,1	453.657,8	1	21.674	149.900	407.143	1.744.062
AM3	341.775,9	445.486,5	0	23.036	149.900	407.143	1.744.062
AS1	346.414,6	430.083,4	0	43.081	150.152	455.234	1.744.062
AS2	334.731,5	418.901,6	0	42.673	149.900	407.143	1.744.062
AS3	340.831,0	429.669,7	0	42.735	149.900	407.143	1.744.062

frequency in the speech transcripts is again clearly non-normally distributed. No difference is found between people with aphasia and controls at any testing moment. A Kruskal-Wallis test did show a difference between the three aphasic test sessions ($X^2(2) = 13,7$, $p = 0,001$). A post-hoc Dunn test reveals that SUBTLEX frequency values of the words used in session 1 are significantly higher than those in session 2 and 3, while session 2 and 3 do not differ between them (1-2: Z = 3,64, $p_{\text{adj}} = 0,001$; 1-3: Z = 2,44, $p_{\text{adj}} = 0,022$). Mild and more severe aphasic participants do not seem to perform differently.

4.4 Discussion

I collected spontaneous speech from four people with mild forms of aphasia, three people with more severe forms of aphasia and seven matched controls. The control participants were recorded once, the aphasic participants were recorded at 2, 5 and 8 months post onset. All participants performed a picture description task, a movie description task and participated in a free conversation, amounting to approximately forty-five minutes of speech per person. Although still small, this corpus is the largest of its kind to my knowledge. Word frequencies were examined for all transcripts to see if Zipf's law holds, both in the form of Zipf-Mandelbrots law and Zipf's β-law. Age of acquisition norms and values for general frequency in Dutch (via SUBTLEX; all norms acquired through the Dutch Lexicon Project 2, Brysbaert, Stevens, Mandera & Keuleers, 2016) were collected for those words for which they were available.

Impairments of the language system in the form of non-fluent aphasia seem to have no effect on the distribution of word frequencies in speech. No differences between control participants and the combined groups of aphasics were found: both for Zipf-Mandelbrot's law and for Zipf's β-law it was found that the values of the parameters were the same for both groups at all testing moments. Fit was high in all cases too, without any meaningful differences between groups.

However, it seems to be the case that this lack of difference is due to the mild nature of the aphasia of most participants. Comparing the more and less severely affected participants with each other shows that the values of ZM-α are higher for the participants with more severe forms of aphasia, without any overlap between the two groups. This result is mirrored in the values that were found for Z-β, which were in all cases lower for the speakers with more severe forms of aphasia when compared to the participants with mild or no aphasia. The lack of overlap between groups is striking and suggests that there might indeed be a difference between more and less severely affected aphasic speakers. No such difference was found for fit or ZM-β, for which the split according to severity does not change the picture in any way. No difference was found across testing sessions. Unfortunately, groups split for severity were too small to make any statistical comparisons. Further research is necessary to see if this difference holds if larger groups are examined.

A steeper slope to Zipf-Mandelbrots law indicates that frequent words are used even more frequently, and infrequent words are used even more infrequently. In other words, infrequent words are impaired to a disproportionally large extent. The finding that people with aphasia struggle significantly more than healthy control participants when the required amount of resources is high (also in non-language related tasks) is in line with work by Van Ewijk (2013). Van Ewijk presented both aphasic and healthy control participants with visual displays of circles. Participants were asked to indicate if any of those circles was smaller than the others. She found that both aphasic and control participants responded slower when more circles were involved and when the size difference was smaller, but the difference in response times was much larger for the aphasic participants than for the control participants. The aphasic participants were thus much more affected by the increased complexity than the control participants. The difference in slope of Zipf's law between the more and less severely aphasic participants could be a result of the same effect. Both healthy and aphasic speakers will use difficult words less frequently, but this process is amplified in aphasia.

The general shape of the distribution of Zipf's law, however, remains unchanged. The downward curvature for high frequency words is not any different for the two groups (as reflected by ZM-β) and neither is the fit of the model (R^2). This suggests that generally speaking, people with aphasia store and retrieve words in the same way as people without aphasia.[20]

The calculation of the Normalised Frequency Difference (NFD) that was developed by Bentz and colleagues did not add anything to the picture. In fact, it was found that the difference within the control transcripts, between the first and second half of the texts, were larger than the difference between the first and second testing session of the aphasic participants and between the controls and third session of the aphasic participants. The NFD can be of value when comparing written texts from different languages (as Bentz and colleagues (2017) did) but seems to be unsuitable to compare spoken texts from the same language.

For AoA, it was found that mean AoA was lower for people with aphasia compared to those without aphasia, meaning that on average, the words that they used are acquired earlier in childhood. This difference was larger for the participants with more severe forms of aphasia. This confirms the hypothesis that words that were acquired earlier in life are better retained in case of damage to the system of language production. No difference was found between testing sessions.

The results for AoA are different from the results for SUBTLEX frequency. This suggests that they are indeed two separate effects: if the AoA effect were a confound of the frequency effect (or the other way round) then no such differences would be expected. For SUBTLEX frequency, no difference between people with and without aphasia was found at any testing session. However, there was a difference between testing sessions within the group of people with aphasia: frequency values of the words used in session 1 were found to be significantly higher than those in session 2 and 3. This difference persists both for mild and more severe aphasic speakers. This difference might be the result of recovery, since high frequency words in the mental lexicon can be considered to be closer to the threshold of lexical activation. Words with a high frequency are generally considered easier to process (e.g. Kittredge et al., 2008, see also Chapter 2, Section 2.1.2). However, the absence of any other effects of recovery in this study renders this hypothesis somewhat unlikely. It might also be due to the nature of the conversations held at the different testing sessions. In session 1, conversations centred around their stroke narratives, while in session 2, family life was often main topic of conversation. However, it is unclear why stroke narratives would elicit more high frequency words than stories about family life. In

[20] This might not seem to be of much interest, but for people with aphasia, it is important. It means that they did not lose part of their knowledge. For their self-esteem, this means everything (as one of them explained to me).

any way, it seems safe to conclude that people with aphasia do not perform differently from healthy speakers when general frequency in Dutch is concerned.

Combined, the results for AoA and SUBTLEX frequency are in line with previous literature, in which it is generally found that the effect of AoA on language processing is stronger than the effect of frequency. This could mean that AoA is of larger influence on complexity than general frequency in the language under consideration.

Another important and for speech therapy potentially relevant conclusion is that the patients that were included in this study were being treated by different speech therapists from different rehabilitation centres. In addition, the nature of their aphasia was for all slightly different, since after all aphasia is a very diverse condition. As a result, they all received different kinds of therapy. Nevertheless, they performed rather uniformly on the measures performed here: there were no aphasic speakers with aberrant parameter values. This study looked at changes in the aphasic lexicon during recovery, not the way in which these changes came to be. That means that it was not a requirement of this study to control for the kind of therapy that patients received. Therapy might be of some influence on the results but is not a threat to it: it is the reality of how actual patients recover.

4.5 Conclusion

Speech from people with aphasia conforms to Zipf's law in the same way as speech from people without aphasia. No differences in parameter values or fit to Zipf-Mandelbrots law or Zipf's β-law were found at any of the testing moments, although there might be an effect of severity on the slope of the distribution: the slope of Zipf-Mandelbrots law seems to be somewhat steeper for more severely aphasic speakers, while the slope of Zipf's β-law seems to be somewhat shallower. The general shape of the distributions did not differ among groups. This finding might reflect the fact that people with aphasia are disproportionally affected when complexity is increased. People with aphasia use on average words that were acquired earlier in life (earlier AoA), while general frequency in Dutch (SUBTLEX frequency) has no influence. This confirms the hypothesis that AoA is of larger influence on word complexity than SUBTLEX frequency.

5 Zipf’s law in aphasia cross-linguistically[21]

In the previous chapter, it was shown that in Dutch, speech from people with non-fluent aphasia conforms to Zipf's law in the same way as speech from people without aphasia, both in slope (ZM-α) and in fit (ZM-R^2). The current chapter extends these findings to other languages, namely to English, Greek and Hungarian. The combination of these languages allows for a comparison between the morphologically poor languages Dutch and English on the one hand and the morphologically rich languages Greek and Hungarian on the other hand. In contrast to the results from the Dutch aphasic speakers it is found that English, Greek and Hungarian aphasics do display a different slope of their word frequency distribution. In addition, this chapter not only considers speech from non-fluent aphasic speakers, but also from the less studied group of fluent aphasic speakers. It will be shown that despite their differing impairments, both groups of aphasics are indistinguishable in terms of Zipf´s law.

5.1 Introduction

As discussed in Chapter 2 and (more briefly) in Chapter 4, aphasia can be divided into several sub-types, depending on the exact impairments of the patient. The broadest distinction that is being made is the one between fluent and non-fluent aphasia. Non-fluent aphasic speech is usually described as effortful and telegraphic, with multiple omissions and/or substitutions, mostly of functional categories (Goodglass, Fodor, & Schulhoff, 1967). As a result, non-fluent aphasic speech sounds markedly different from healthy speech but is still meaningful. On the contrary, fluent aphasic speech usually sounds effortless and continuous. Nevertheless, it is often hard to follow: paraphasias and neologisms are included

[21] The aphasiabank results were previously published as Neophytou, K., Egmond, M. van & Avrutin, S. (2017). Zipf's law in aphasia across languages: A comparison of English, Hungarian and Greek. *Journal of Quantitative Linguistics*, 24(2-3), pp. 178-196.

Terminology on Zipf's law

The terminology used in this chapter is the same as elsewhere in this dissertation, but is repeated here for convenience and in an attempt to limit confusion to a minimum (see also Chapter 2, section 2.2.1).
The term ‘Zipf's law’ is used as an overarching term for two formulas:

Zipf-Mandelbrot’s law:

7) $f(w) = \frac{C}{(r(w) + \beta)^{\alpha}}$

where the frequency *f* of word *w* is determined by its rank *r* when all words are ordered from most to least frequent, the parameters *α* and *β*, and a text size dependent constant *C*, and

Zipf’s *β*-law:

8) $n_f = C \cdot f^{-\beta}$

where the number of words *n* with frequency *f*, in other words, the size of the frequency class, is determined by parameter *β* and a text size dependent constant *C*.

Both Zipf-Mandelbrot’s law and Zipf’s *β*-law are formulations of the same phenomenon, and the unspecific term ‘Zipf's law’ has throughout the literature been used to describe both. To avoid confusion, when either of the two laws is discussed specifically it will be called by its full name.
Another potential point of confusion is the parameter *β*, because both laws have one. Therefore, the *β* from Zipf-Mandelbrot’s law will be dubbed ZM-*β*, while the *β* from Zipf’s *β*-law will be dubbed Z-*β*.

while content words are omitted or substituted (Bastiaanse, 2011; Wernicke, 1874). This often results in incomprehensible phrases and sentences. The difficulties of non-fluent aphasic people in speech production are thought to be due to the reduction of processing resources, not to the loss of syntactic abilities (Avrutin, 2006, see also Chapter 2, section 2.1.3). Previous studies show that the syntactic processes are intact in people with non-fluent aphasia (e.g. Zurif, Swinney, Prather, Solomon, & Bushell, 1993). The problem is that they are slower. The brain injuries in these patients affect the amount of available processing resources. In order to compensate for the reduction of those resources, more processing time is required, thus causing the various distortions in speech.

As with non-fluent aphasia, the language problems of fluent aphasics are also thought to relate to the processing system and not to the lexicon itself, but they surface very differently. The content words that fluent aphasics have problems with are mostly finite verbs (Bastiaanse, 2011). Specifically, the variability in finite verbs is reduced compared to the variability in infinitive verbs. This shows that there is no lexical retrieval problem in non-fluent aphasics, but either a grammatical complexity problem or an integration problem: a problem that is encountered when the lexical form and inflection for tense and agreement need to be integrated. Since, similarly to non-fluent aphasia, the problem is not in the lexicon itself, the processing system is likely to be responsible for the linguistic difficulties of fluent aphasic people as well.

As discussed in Chapter 4, Section 4.1.1, Zipf's law in aphasic speech has only little been studied. Previous work reported in Van Egmond, Van Ewijk and Avrutin (2015; see also Van Ewijk, 2013) was the first study into Zipf's law in aphasia. We interviewed four participants with chronic non-fluent aphasia and compared their speech to that of four healthy gender and age matched controls from the Corpus of Spoken Dutch (CGN-consortium, 2014). We found no difference in fit to the traditional Zipf's law for both groups: both had a very good fit above $R^2 = 0,95$. We did find a difference in slope: for aphasics, a value of $\alpha = 0,834$ was found, while for controls $\alpha = 0,677$ was found, which reflects the reduced variability in the vocabulary of the aphasic speakers. Similar to the current data set, the speech samples were very short, only 352 tokens long.
The results from Van Egmond et al. are different from those of the aphasic speakers in the previous chapter. In the study reported on in Chapter 4 I tested seven non-fluent aphasic speakers who were 2, 5 or 8 months post onset (which is considered the post-acute stage of recovery). For this group, no difference in fit or in slope was found, although it seemed to be the case that a difference in fit might exist for the more severely aphasic speakers (the small sample size hindered statistical testing of this difference).

The aim of the current chapter is twofold. First, it is tested to what extent the non-fluent aphasic speech findings from the previous chapter and from previous work reported on in Van Egmond, Van Ewijk and Avrutin (2015; see also Van Ewijk, 2013) can be generalized to other languages. It is tested whether speech from aphasic speakers in English, Hungarian and Greek conforms to Zipf's law too, and if so, if there are any differences in slope.

Second, this chapter reports on Zipf 's law in fluent aphasic speech. To the best of my knowledge his has not been done before. It is expected that the paraphasias and neologisms in fluent aphasic speech contribute to a shallower slope of Zipf-Mandelbrot's law compared to healthy speech. These utterances are likely to be uttered only once or twice, thus creating extra hapax legomena. A large number of hapax legomena results in a shallower slope.

In Chapter 3, it was concluded that at least 1500 tokens are necessary for the study of Zipf's law. However, sometimes, longer samples are not available, like in the case of spoken language from an impaired population as studied here. The data that are being used come from the AphasiaBank corpus, a corpus of cross-linguistic aphasic speech. It would be a waste of resources if this valuable source of data could not be used, since obtaining such a corpus requires an enormous amount of time and effort. This chapter therefore proposes a solution to this problem in the form of comparisons to multiple same-sized samples from baseline corpora of comparable speech from a healthy population. Through this method, parameter values from the aphasic speakers are validated by retrieving the parameter values of a large number of equally sized samples from corpora of spontaneous speech in the respective languages. These data then function as a baseline, aiding in the interpretation of the data from the aphasic speakers.

Because of the small samples, only Zipf-Mandelbrot's law is used here, not Zipf's β-law. In Chapter 3, section 3.4 it was discussed that Zipf's β-law is more sensitive to disruptions of the frequency distribution than Zipf-Mandelbrot's law. However, a downside of this formulation of Zipf's law is that text samples need to be somewhat larger to prevent empty frequency classes. The samples used here (200 tokens) do not fit this criterium and are thus too small for Zipf's β-law to be properly studied.

Three analyses were performed. First, the findings from the aphasic speakers are compared with the outcomes from the baseline corpora of unimpaired speech. If variation within a large corpus of healthy speech is so large that it incorporates the values found for aphasic speakers, then it is questionable to what extent any differences between the two groups are meaningful. If, on the other hand, the values for the aphasic speakers fall outside of the normal range determined by the baseline corpus, then this is extra evidence that any difference is meaningful and real.

Next, a comparison between the frequency distributions of healthy and aphasic speakers in a language other than Dutch is given, namely in English. Based on the previous chapter and on Van Egmond et al. (2015), no differences in fit are expected: all frequency distributions are expected to conform to Zipf-Mandelbrot's law. A difference in slope (ZM-α) might or might not exist, but if it does, non-fluent aphasic speakers are expected to have a steeper slope (reflected in a higher ZM-α). As discussed above, fluent aphasic speakers are expected to have a shallower slope (lower ZM-α).

Finally, the properties of Zipf's law in aphasic speech are studied across different languages, both in fluent and non-fluent aphasia. More specifically, speech samples in Greek, English and Hungarian are examined. In this between-languages analysis, similarly to the within-languages analysis, no differences in the fit to Zipf's law are expected, but it is expected that differences exist in slope between the different languages.

The morphology of a language, specifically the extent to which it applies inflection and its types of word formation, are known to be of influence on the parameters of Zipf's law. Bentz, Kiela, Hill, and Buttery (2014) found a difference in the parameters of Zipf's law between Old English and Modern English, a difference that lies in the morphological variability of these two historical versions of English. Specifically, a language with poor morphology, like Modern English, displays a steeper slope than a language with rich morphology, like Old English. Here, it is investigated whether fluent and non-fluent aphasic speech also reflects this difference in morphological variability among the different languages that are tested. Given that Zipf's law is intact in aphasic speech, it is expected that speech from the morphologically simple English has a steeper slope compared to speech from the morphologically complex Hungarian and Greek.

5.2 Methods

As the parameters of Zipf's law are highly dependent on text size, the same number of tokens was selected for each speaker. In all cases, the first 200 words were analyzed. This number formed the balance between both including as many speech samples as possible and including as many words as possible, because many participants' recordings were relatively short.

This is a very short sample size, however. In fact, it is too short, as it was argued in Chapter 3 that a minimum of 1500 tokens is required for calculations of Zipf's law. Unfortunately, longer samples were not available. As was discussed in the introduction, a baseline was established to be able to say something about any potential differences between these samples nonetheless. This baseline consists of a large number of 200-word samples from an independent spoken language corpus in the language under consideration. Such a baseline allows us to answer the question how much variation is to be expected for such small samples, and also how the values that are found relate to the values for the language in general. The variation found in the spontaneous speech samples can thus be interpreted in light of this baseline. Zipf's law is language dependent, so a baseline is established for each language separately.

Theoretically, this method allows for a baseline tailored to the length of each speech sample individually, thus allowing to use its full size. The downside of this is that in that case, no comparisons between groups are possible based on the parameter values. For that reason, it was decided to maintain equal sample sizes per speaker.

First, the methods concerning these baseline corpora will be discussed, followed by the methods concerning the aphasic speech samples. The methods applied for the calculations of Zipf's law are identical for all samples and discussed last.

5.2.1 Baseline corpora

English

Data used to calculate the baseline for English has been extracted from the British National Corpus (BNC), managed by Oxford University Computing Services on behalf of the BNC Consortium. The BNC is a 100-million-word collection of samples of written (90%) and spoken (10%) language from a wide range of sources, designed to represent a wide cross-section of British English, both spoken and written, from the late twentieth century. The spoken language was collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins. Work on building the corpus began in 1991, and was completed in 1994.

Only texts consisting of more than 200 tokens that were originally spoken were selected, and from this only those texts were selected that were either specified as recorded in leisure context or as recorded conversations. These texts were split per speaker, and only those files that contained at least 200 tokens were kept for further analysis.

Greek

The material used to calculate the baseline for Greek came from the Corpus of Spoken Greek, in full the Spoken Language Corpus of the Institute for Modern Greek Studies (Manolis Triandafyllides Foundation), made available with the consent of Professor Th.-S. Pavlidou, director of the research program "Language interaction and conversational analysis". It utilizes earlier data collections from various earlier research and/or student projects since the 1980s. The material has been drawn from naturally-occurring circumstances of communication and comprises of everyday conversations among friends and relatives, telephone calls, classroom interaction, television news, and other television broadcasts. It was originally designed for the qualitative analysis of language and linguistic communication, especially from the perspective of Conversation Analysis.

Two transcripts were made available to me, both consisting of everyday conversations among two non-overlapping groups of four friends and/or relatives. Files were split per speaker for further analysis.

Hungarian

The material used to calculate the baseline for Hungarian was taken from the Budapesti Szociolingvisztikai Interjú (Budapest Sociolinguistic Interview, BSI), owned by the Hungarian Research Institute for Linguistics, Hungarian Academy of Sciences (RIL HAS).

BSI-2 was compiled in 1987 and contains speech from 50 participants: 10 teachers of over 50 years of age, 10 university students, 10 sales clerks, 10 factory workers and 10 high school students (around age 16). Each interview consists of a test and a guided conversation, about 90 minutes for each informant. Here, only the guided conversations were used.

5.2.2 AphasiaBank samples

All spontaneous speech transcripts from people with aphasia were taken from AphasiaBank (MacWhinney, Fromm, Forbes, & Holland, 2011), an online database with speech from aphasic patients in different languages. All texts were selected that fitted two criteria: the speech had to be spontaneous and non-directed (i.e. an unstructured interview) and the patient's aphasia type should be clearly categorized into fluent or non-fluent. This resulted in English transcripts from fluent aphasic, non-fluent aphasic and healthy speakers; and Greek and Hungarian transcripts from fluent and non-fluent aphasic speakers (English: Bates, Friederici, Wulfeck, & Juarez, 1988; Hungarian: MacWhinney & Osmán-Sági, 1991; Greek: Goutsos, Potagas, Kasselimis, Varkanitsa, & Evdokimidis, 2011). No healthy speech samples were included in the corpus for Greek, whereas the healthy Hungarian speech samples that were included in the corpus were collected through directed speech tasks, and therefore did not meet our data inclusion criteria. Detailed information for the participants can be found in Table 5.1 (English), Table 5.2 (Greek) and Table 5.3 (Hungarian).

All conversations were in CHAT-format (MacWhinney, 2000). In the analysis for each conversation, the speech from the investigator was ignored.

Table 5.1. Speaker details: English speakers (time after injury was not provided).

	Speaker ID*	**Age**	**Sex**	**Type of Aphasia**
Fluent Aphasics	ACWT11a	48;4	male	Wernicke
	adler23a	81;3.	male	Wernicke
	BU10a	75;9.	male	Wernicke
	elman12a	57;4.	male	Wernicke
	kansas14a	77;4.	female	Wernicke
	kansas23a	75;6.	female	Wernicke
	kurland01c	58;9.	male	Wernicke
	kurland01d	59;0.	male	Wernicke
	scale11b	91;8.	female	Wernicke
	scale24a	61;9.	male	Wernicke
	tucson03a	46;9.	female	Wernicke
	tucson13a	68;2.	male	Wernicke
	whiteside10a	66;2.	male	Wernicke
	williamson23a	60;9.	male	Wernicke
	garrett01a	76;8.	female	Wernicke
	kansas05a	69;10.	female	Wernicke
	thompson03a	67;6.	male	Wernicke
	thompson05a	63;10.	female	Wernicke
	tucson15a	74;1.	male	Wernicke
	elman14a	76;3.	female	Wernicke
	whiteside14a	40;6.	female	Wernicke
	ACWT10a	48;4.	male	Wernicke
	adler06a	70;7.	male	Wernicke
	kansas12a	N/A	male	Wernicke
Non-Fluent Aphasics	adler13a	52;4.	male	Broca
	adler16a	63;6.	male	Broca
	adler25a	66;2.	male	Broca
	BU07a	52;4.	male	Broca
	BU08a	*64;6.*	*male*	Broca
	elman03a	*64;6.*	*male*	Broca
	elman06a	76;10.	female	Broca
	elman11a	52;1.	male	Broca
	fridriksson03a	46;3.	female	Broca
	fridriksson10a	64;9.	male	Broca
	fridriksson12a	47;10.	female	Broca
	kempler03a	64;6.	male	Broca
	kempler04a	60;3.	female	Broca
	kurland10b	78;4.	female	Broca
	scale01a	78;3.	male	Broca
	scale15b	59;4.	male	Broca
	scale25a	52;7.	female	Broca
	scale26a	58;9.	male	Broca
	tap13a	49;3.	female	Broca
	tap19a	N/A	female	Broca
	tcu02a	42;7.	male	Broca
	tcu03a	41;9.	male	Broca
	tcu07a	49;2.	female	Broca
	tcu08a	57;2.	male	Broca
	whiteside15a	53;10.	female	Broca
	wright206a	39;0.	female	Broca

* Names in SpeakerID labels refer to subcorpora, not to participants.

Table 5.1 (continuation). Speaker details: English speakers

	Speaker ID*	**Age**	**Sex**	**Type of Aphasia**
Healthy Controls	capilouto02a	85;2.	male	.
	capilouto03a	75;0.	female	.
	capilouto04a	80;6.	female	.
	capilouto05a	72;3.	male	.
	capilouto06a	82;4.	female	.
	capilouto07a	72;0.	female	.
	capilouto08a	74;0.	male	.
	capilouto09a	82;7.	male	.
	capilouto10a	72;11.	male	.
	capilouto11a	53;5.	male	.
	capilouto12a	54;11.	female	.
	capilouto13a	71;5.	female	.
	capilouto14a	81;1.	male	.
	capilouto15a	71;10.	male	.
	capilouto16a	79;11.	female	.
	capilouto17a	71;3.	female	.
	capilouto18a	64;4.	female	.
	capilouto19a	60;9.	male	.
	capilouto20a	71;6.	female	.
	capilouto21a	74;6.	male	.
	capilouto23a	70;6.	male	.
	capilouto24a	70;8.	male	.
	capilouto26a	77;0.	male	.
	capilouto28a	76;9.	male	.
	capilouto29a	71;6.	male	.
	capilouto31a	72;2.	female	.

* Names in SpeakerID labels refer to subcorpora, not to participants.

Table 5.2. Speaker details: Greek speakers

	Speaker ID	**Age**	**Sex**	**Type of Aphasia***	**Time after injury**
Fluent Aphasics	02_combined	71	female	Fluent	17 months
	04_combined	78	female	Fluent	3.5 years
	15_combined	72	male	Fluent	8 days
	19_combined	64	female	Fluent	9 days
	20_combined	74	female	Fluent	4 months
	30_combined	60	female	Fluent	20 days
	32_combined	73	male	Fluent	20 days
	33_combined	34	female	Fluent	8 days
	34_combined	68	male	Fluent	4.5 years
	37_combined	58	female	Fluent	2.5 months
	38_combined	72	female	Fluent	3.5 years
	39_combined	55	female	Fluent	20 days
Non-Fluent Aphasics	05_combined	58	male	Non-fluent	4 weeks
	09_combined	56	male	Non-fluent	3 months
	11_combined	50	male	Non-fluent	20 months
	35_combined	86	female	Non-fluent	2 days
	36_combined	63	male	Non-fluent	6 days

* The compilers of the corpus only used the labels Fluent and Non-fluent to describe the type of aphasia for the Greek aphasic groups

Table 5.3. Speaker details: Hungarian speakers

	Speaker ID	Age	Sex	Type of Aphasia	Time after injury
Fluent Aphasics	c01_combined	54	male	Conduction	9.5 months
	c02_combined	35	female	Conduction	4 months
	c05_combined	52	male	Conduction	3 months
	w02_combined	51	female	Wernicke's	7 months
	w04_combined	55	female	Wernicke's	4 months
Non-Fluent Aphasics	b05_combined	44	male	Broca's	7.5 moths
	b07_combined	57	female	Broca's	21 months
	a06_combined	33	male	Anomia	10 weeks
	a07_combined	64	male	Anomia	10 months
	a11_combined	59	male	Anomia	6 months

5.2.3 General methods

Text preparation

All symbols, markers and punctuation markers were ignored, leaving only the bare words for further analysis. All characters were converted to lower-case. From each baseline corpus file, samples of 200 tokens were extracted. Samples were taken at random, maximizing the total number of samples without allowing for overlap. For the AphasiaBank files, in all cases the first 200 tokens were selected for further analysis.

Analysis

The formula that is used for the Zipf's law-calculations is Zipf-Mandelbrot's law, provided as Formula 1 in the box at the beginning of this chapter. Both ZM-α and ZM-β (Mandelbrot's β) were calculated. The values of the parameters were calculated through maximum likelihood estimation (Murphy, 2015), after making a first approximation of the parameters through linear regression (Izsák, 2006).

A combination of the R^2-values and a rating of the parameter values is used to determine whether Zipf-Mandelbrot's law holds. Although it is still a matter of debate of when it can be claimed that Zipf-Mandelbrot's law does not hold, the closer the R^2-values are to 1, the better the fit to Zipf-Mandelbrot's law is. But this only holds if the parameter values are within the normal range, meaning that ZM-β is no larger than 10 and ZM-α is somewhere between 0 and 2. These values were also statistically compared to see if any of the groups displayed a significantly lower fit than the others. The alpha values are statistically compared between groups. The ZM-*β*-values were calculated but not further analyzed. In these samples, the number of ranks on which this value is based is so small that statistical analyses are unreliable.

Comparisons with the baseline corpora

The question to answer is if the AphasiaBank samples are likely to come from the same population as the baseline corpora samples. The null hypothesis is that both fluent and non-fluent aphasic and healthy speakers come from the same population as the speakers in the baseline corpora, which means that their parameter values are indistinguishable. The alternative hypothesis is that (any of) the AphasiaBank samples come from a different population. For the controls, such a difference would be surprising, most likely caused by a difference in test methods. In that case, the results from the people with aphasia should also be treated with caution. If the control speakers are from the same population but the aphasic speakers are not, then this is a clear sign that their language disorder is the reason for that.

The method applied is permutation tests. This test and not the traditional ANOVA was chosen because of the large difference in number of samples from the baseline corpora and from the AphasiaBank samples, rendering ANOVAs unsuitable. A permutation test makes use of resampling. It tests how likely is it that samples come from the same population by randomly redrawing samples (with the same size as the original groups) from the combined set of values of both the baseline corpus and the AphasiaBank samples. The mean is calculated for each redraw. It is then determined how often the difference in mean is larger than that of the original groups. If the difference is larger in less than 5% of the cases then the difference is considered significant.
Using this method, we can interpret the differences found within the AphasiaBank corpus in light of the variation within a much larger corpus.

Group wise analyses (comparable to traditional ANOVAs) are performed using the R-function 'independence_test' from the package 'coin'. Post hoc testing is done using the function 'pairwisePermutationTest' from the package 'rcompanion'. This test uses the Benjamini and Hochberg (1995) correction to multiple testing. This method controls the false discovery rate, the expected proportion of false discoveries amongst the rejected hypothesis. This condition is less stringent than the family-wise error rate used by for example the well-known Bonferroni correction, rendering it a powerful method (R Core Team, 2017).

Comparisons within the AphasiaBank samples

Comparisons within the AphasiaBank samples were performed if this made any sense in light of the comparison with the baseline corpora.
For the within-language analysis, the comparisons were made between fluent, non-fluent and healthy speech for English only. For the between-languages analysis, the comparisons were made between fluent aphasic speech and non-fluent aphasic speech for all three languages.
For the within-language analysis, a simple ANOVA analysis was performed since there was only a single independent variable: group. For the between-languages analysis, a factorial ANOVA was conducted because there were two independent variables: group and language.

5.3 Results

5.3.1 Corpus descriptives

English baseline corpus

The text samples were highly diverse, recorded at a multitude of circumstances and situations. This resulted in a number of obvious outliers, some even reaching the artificial boundary of ZM-α = 10 which was the search limit for the software. It was decided to remove all samples for which ZM-α was more than 3 standard deviations above the mean (more than 3 standard deviations below the mean did not occur). These outliers generally contained a very high word repetition rate within the text and could not be considered normal speech in conversation (for example recordings in which someone was counting or listing something). This procedure resulted in the removal of 0,81% of all samples (225 samples from 61 different speakers).

Descriptives of the remaining corpus can be found in Table 5.4. The parameter distributions are plotted in Figure 5.1.

The selected samples come from 1.832 speakers in 1.689 conversations (601 females, 721 males, 516 unspecified). Age was not specified for 877 speakers. The other speakers were rather evenly distributed over the used age groups: 158 speakers had age 0-14 years, 161 had age 15-24 years, 171 had age 25-34 years, 140 had age 35-44 years, 151 had age 45-59 years and 180 had age 60+ years. 749 speakers had no specified dialect, the others were specified as having one of 28 different dialects.

Table 5.4. Descriptives of the English baseline corpus BCN

N samples	27.492
N speakers	1.838

	ZM-*α*	**ZM-*β***	**ZM-R²**	**TTR**
Min	0,391	-1,000	0,822	0,180
1st Qu.	0,642	0,374	0,955	0,495
Median	0,709	1,210	0,964	0,530
Mean	0,724	1,675	0,962	0,528
3rd Qu.	0,788	2,424	0,972	0,560
Max.	1,3310	26,958	1,000	0,770
SD	0,121	1,923	0,014	0,052

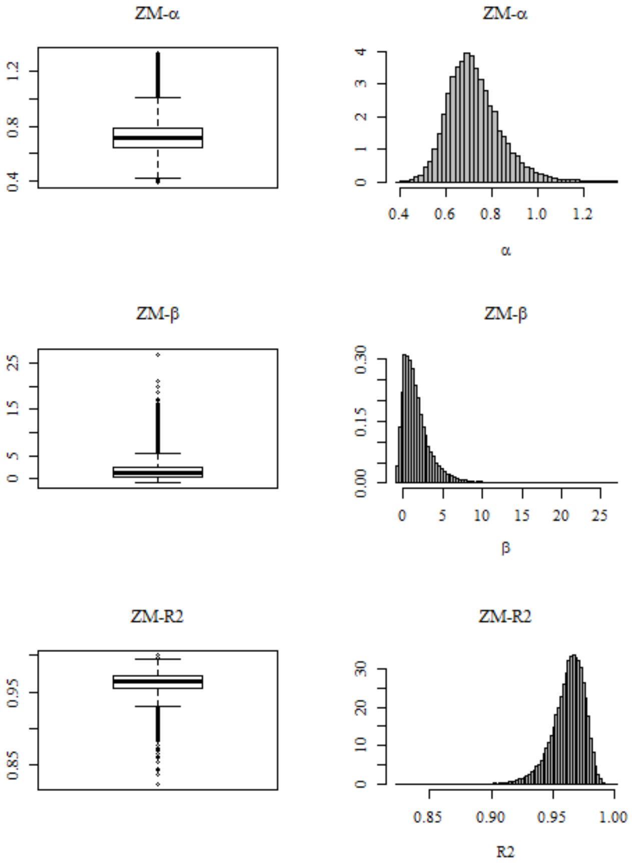


Figure 5.1. Parameter distribution for the English baseline corpus BCN

Greek baseline corpus

Total type and token counts for the eight individual speakers are given in Table 5.5. Sampling sections of 200 tokens resulted in a total of 65 samples, divided over the eight speakers. No outliers were present. The descriptives of all samples are given in Table 5.6. The parameter distributions are plotted in Figure 5.2.

Table 5.5. Types and tokens for the Greek baseline corpus per speaker

Text	**Speaker**	**N Types**	**N Tokens**	**TTR**
5-005 I.22.A.3	Φωτεινή	619	1354	0,457
5-005 I.22.A.3	Χρυσή	730	1768	0,413
5-005 I.22.A.3	Μυρτώ	778	2008	0,387
c5-005 I.22.A.3	Άρτεμη	738	1795	0,411
6-006 I.22.A.3.2	Λίνος	812	2101	0,386
6-006 I.22.A.3.2	Ρόζα	678	1443	0,470
6-006 I.22.A.3.2	Μάρα	869	2235	0,389
6-006 I.22.A.3.2	Στάθης	591	1091	0,542
	Total	*3245*	*13.795*	*0,235*

Table 5.6. Descriptives of the Greek baseline corpus

N samples	**65**
N speakers	**8**

	ZM-*α*	**ZM-*β***	**ZM-R^2**	**TTR**
Min	0,358	-0,878	0,871	0,475
1st Qu.	0,446	-0329	0,916	0,631
Median	0,485	0,103	0,936	0,675
Mean	0,492	0,465	0,932	0,656
3rd Qu.	0,534	0,668	0,947	0,704
Max.	0,671	4,403	0,971	0,780
SD	0,067	1,185	0,020	0,069

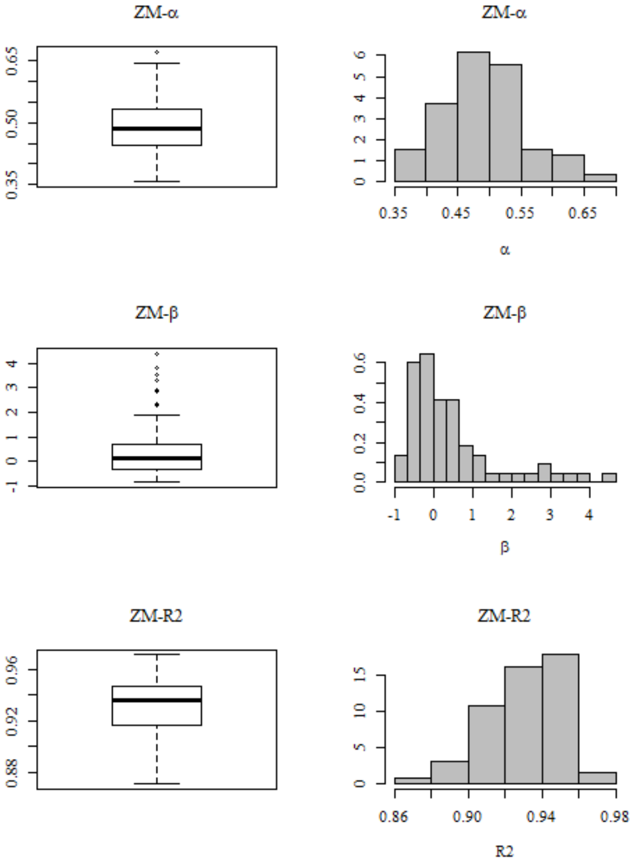


Figure 5.2. Parameter distribution for the Greek baseline corpus

Hungarian baseline corpus

Thanks to the methodology of similarly structured guided interviews, no outliers were present. A total of 832 samples could be extracted, divided over the 50 speakers. Descriptives of the corpus can be found in Table 5.7. The distribution of the parameters is plotted in Figure 5.3.

Table 5.7. Descriptives of the Hungarian baseline corpus

N samples	**832**
N speakers	**50**

	ZM-*α*	**ZM-*β***	**ZM-R^2**	**TTR**
Min	0,380	-0,986	0,804	0,455
1st Qu.	0,511	-0,548	0,927	0,600
Median	0,556	-0,244	0,945	0,635
Mean	0,561	-0,130	0,941	0,634
3rd Qu.	0,610	0,140	0,960	0,665
Max.	0,800	2,728	0,992	0,765
SD	0,071	0,572	0,026	0,048

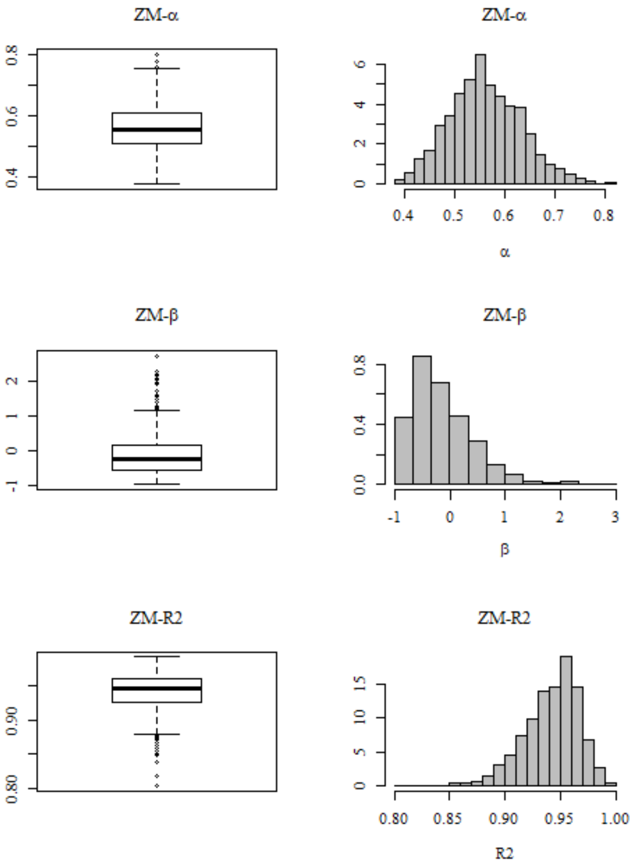


Figure 5.3. Parameter distribution for the Hungarian baseline corpus

AphasiaBank samples

Details about number of samples and mean number of tokens in the 200-word samples are given in Table 5.8. Details about the parameter values and type/token ratios are given in Table 5.9. Zipf-Mandelbrot's law for all AphasiaBank samples per group is given in Figure 5.4. A boxplot depicting the outcomes for ZM-*α* for the three groups in English is given in Figure 5.5. A boxplot depicting the outcomes for ZM-*α* for the two groups in English, Greek and Hungarian is given in Figure 5.6.

Table 5.8. Descriptives of the AphasiaBank samples for the 200-word samples

		N	Mean N tokens
English	**controls**	26	99,58
	fluent	24	88,17
	non-fluent	26	76,62
Greek	**fluent**	12	110,67
	non-fluent	5	106,00
Hungarian	**fluent**	5	100,80
	non-fluent	5	102,60

Table 5.9. Parameter values of the AphasiaBank samples

ZM-*α*		**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**SD**
English	**controls**	0,570	0,641	0,660	0,685	0,717	0,976	0,095
	fluent	0,582	0,672	0,775	0,801	0,889	1,162	0,164
	non-fluent	0,593	0,776	0,927	0,928	0,993	1,993	0,276
Greek	**fluent**	0,563	0,606	0,666	0,662	0,701	0,808	0,074
	non-fluent	0,608	0,661	0,728	0,713	0,741	0,826	0,083
Hungarian	**fluent**	0,563	0,701	0,725	0,725	0,794	0,840	0,106
	non-fluent	0,630	0,636	0,647	0,689	0,733	0,798	0,074
ZM-*β*		**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**SD**
English	**controls**	-0,720	-0,431	-0,250	0,144	0,001	4,854	1,287
	fluent	-0,690	-0,037	0,856	1,238	2,078	7,265	1,846
	non-fluent	-0,814	0,234	1,432	1,776	2,082	10,324	2,585
Greek	**fluent**	-0,513	0,219	0,534	0,751	1,414	2,436	0,892
	non-fluent	-0,852	-0,496	-0,244	1,401	1,022	7,578	3,524
Hungarian	**fluent**	-0,824	-0,234	0,388	0,212	0,457	1,275	0,790
	non-fluent	-0,837	-0,552	-0,536	-0,354	0,047	0,110	0,413
ZM-R^2		**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**SD**
English	**controls**	0,897	0,966	0,977	0972	0,984	0,989	0,019
	fluent	0,919	0,966	0,972	0,971	0,982	0,988	0,016
	non-fluent	0,927	0,963	0,973	0,968	0,982	0,990	0,019
Greek	**fluent**	0,933	0,951	0,962	0,960	0,972	0,978	0,016
	non-fluent	0,943	0,946	0,949	0,957	0,959	0,990	0,019
Hungarian	**fluent**	0,942	0,958	0,968	0,964	0,975	0,978	0,015
	non-fluent	0,955	0,956	0,957	0,963	0,971	0,977	0,010
TTR		**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**SD**
English	**controls**	0,420	0,460	0,505	0,498	0,534	0,575	0,044
	fluent	0,310	0,408	0,442	0,441	0,470	0,560	0,056
	non-fluent	0,240	0,351	0,375	0,383	0,428	0,515	0,069
Greek	**fluent**	0,475	0,526	0,558	0,553	0,578	0,635	0,047
	non-fluent	0,505	0,530	0,535	0,530	0,535	0,545	0,015
Hungarian	**fluent**	0,470	0,485	0,485	0,504	0,520	0,560	0,036
	non-fluent	0,455	0,485	0,505	0,513	0,530	0,590	0,051

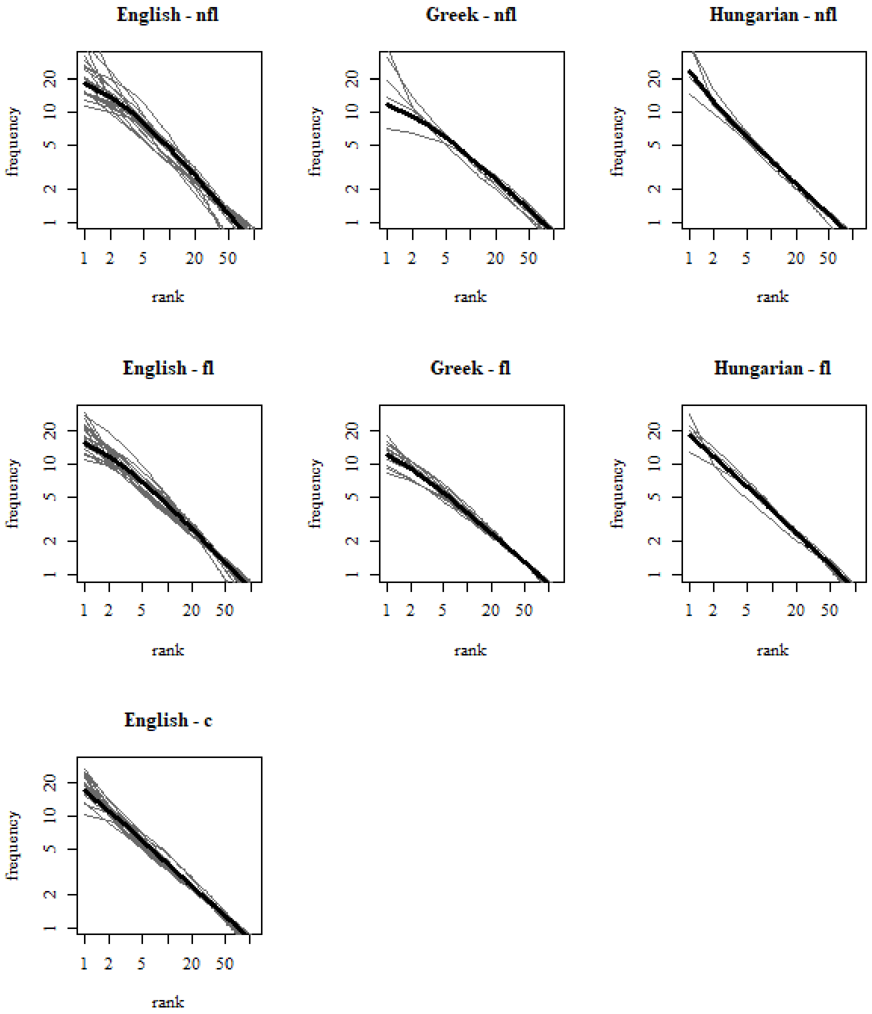


Figure 5.4. Zipf-Mandelbrot's law for all AphasiaBank samples for non-fluent (top row), fluent (middle row) and control speakers (bottom row) in English (left column), Greek (middle column) and Hungarian (right column). Grey lines show the slope for individual texts, black lines show the slope based on mean values per group.

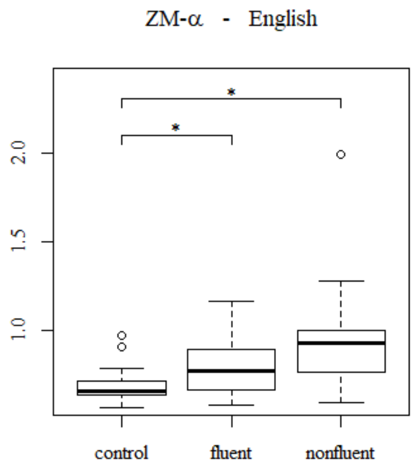


Figure 5.5. Boxplot for ZM-*α* for English control speakers, fluent aphasic speakers and non-fluent aphasic speakers (stars indicate significant differences between groups)

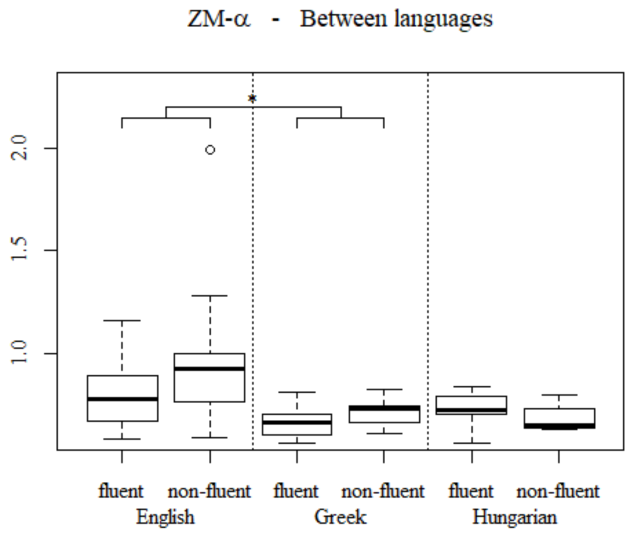


Figure 5.6. Boxplot for ZM-*α* between languages for fluent and non-fluent aphasic speakers (stars indicate significant differences)

5.3.2 Analyses

Baseline corpora vs. AphasiaBank samples

Density plots for ZM-*α* for the baseline corpora and the different groups of the AphasiaBank corpus per language are given in Figure 5.7.

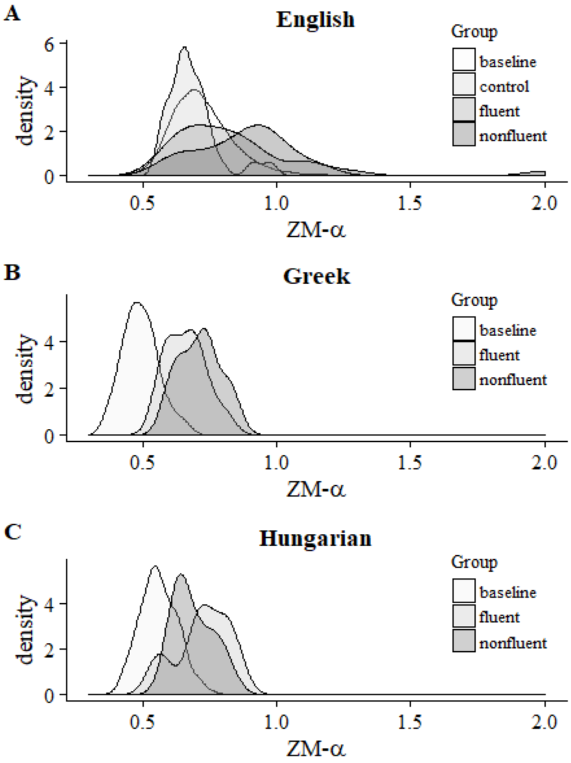


Figure 5.7. Density plots showing the probability distribution of ZM-*α* for the different groups in A English, B Greek and C Hungarian.

English

First, fit is considered. A permutation test shows that a significant difference in R^2-values exists between the baseline corpus and (some of) the AphasiaBank samples (p < 0,001).
Post hoc pairwise permutation tests reveal that this is due to a difference in fit between the baseline corpus and the control speakers and fluent aphasic speakers

(c: p < 0,001; fl: p = 0,008). In both cases, the baseline corpus has the lower R^2. Fit for the non-fluent speakers is not different from the baseline corpus.

For ZM-α, a permutation test shows that a significant difference exists between the baseline corpus and (some of) the AphasiaBank samples (p < 0,001).
Post hoc pairwise permutation tests reveal that both fluent (adj. p = 0,003) and non-fluent (adj. p < 0,001) speakers' samples from AphasiaBank differ significantly from the baseline corpus. The control samples are no different from the baseline corpus samples.

Greek

For fit in Greek, a permutation test shows that a significant difference exists between the baseline corpus and (some of) the AphasiaBank samples (p < 0,001).
Post hoc pairwise permutation tests reveal that this is due to a significant difference between the baseline corpus and both AphasiaBank groups (fl: p < 0,001; nfl: p = 0,01). In both cases, the AphasiaBank samples display the higher fit.

For ZM-α, a permutation test shows that a significant difference exists between the Greek baseline corpus and (some of) the AphasiaBank samples (p < 0,001).
Post hoc pairwise permutation tests reveal that both fluent (adj. p < 0,001) and non-fluent (adj. p < 0,001) speakers' samples from AphasiaBank differ significantly from the baseline corpus. No Greek control speakers were present in the AphasiaBank corpus.

Hungarian

For fit in Hungarian, results are somewhat mixed. A permutation test shows that a significant difference exists between the baseline corpus and (some of) the AphasiaBank samples (p = 0,02).
However, this difference is not found by pairwise permutation tests (specifically, not after the Benjamini and Hochberg (1995) correction to multiple testing): for both comparisons it is found that p = 0,09. The AphasiaBank samples tend to have the higher R^2-values.

For ZM-α, a permutation test shows that a significant difference exists between the Hungarian baseline corpus and (some of) the AphasiaBank samples (p < 0,001).
Post hoc pairwise permutation tests reveal that both fluent (adj. p < 0,001) and non-fluent (adj. p < 0,001) speakers' samples from AphasiaBank differ significantly from the baseline corpus. No Hungarian control speakers were present in the AphasiaBank corpus.

Significant differences exist between the baseline corpora and the AphasiaBank samples for all three languages, while fit is not significantly worse. This means that all samples can be considered to conform to Zipf's law and that it is meaningful to continue with comparisons within AphasiaBank.

Within-language analysis

The within-language analysis concerns the text samples from English fluent and non-fluent aphasic speakers and control speakers.

First, fit is considered. From Levene's test, we are able to conclude that there is homogeneity of variance (F(2, 73) = 0,44, p = 0,43). Using simple ANOVAs, no difference between groups is found. Thus, R^2 is equally high for all groups. The lowest R2 value that is found for any of the texts is R^2 = 0,897, which was for one of the control speakers.

Concerning ZM-*α,* Levene's test shows that there is no homogeneity of variance (F(2, 73) = 4,96, p = 0,04). Reciprocal transformation of the data solves this (F(2, 73) = 2,06, p = 0,14).
Simple ANOVA analysis after data transformation reveals that there exists a significant difference between groups (F(2, 73) = 13,06, p < 0,001). Post hoc tests using the Bonferroni correction show that both the non-fluent group (p < 0,01) and the fluent group (p < 0,05) display a significantly higher ZM-*α* compared to the control group. The aphasic speakers thus display a steeper slope than the healthy controls, reflecting a less varied vocabulary. The fluent group and non-fluent group do not differ significantly from each other.

Between-languages analysis

The between-languages analysis concerns the text samples from English, Hungarian and Greek fluent and non-fluent aphasic speakers. It should be noted though that, as the Greek non-fluent aphasic group and the Hungarian fluent and non-fluent aphasic groups were rather small, statistical analyses are only a tentative indication of the strength of any effect.

Again, first the fit is considered. From Levene's test, it can be concluded that there is homogeneity of variance (F(5, 71) = 0,46, p = 0,81). Using factorial ANOVAs, no difference between groups is found. Thus, R^2 is equally high for all groups. The lowest R^2 value that is found for any of the texts is R^2 = 0,897, which is for one of the English control speakers.

Concerning ZM-α, Levene's test shows that there is homogeneity of variance (F(5, 71) = 1,58, p = 0,18). Factorial ANOVA analysis reveals that there is a significant effect of language (F(5, 71) = 6,19, p = 0,003). There is no significant main effect of group and no interaction effect between language and group. Post hoc tests using the Bonferroni correction show that Greek and English are significantly different from each other (p = 0,003), with English displaying higher ZM-α values (reflecting a less varied vocabulary). The difference between English and Hungarian approaches significance (p = 0,06), with English tending to have higher ZM-α values. No significant difference is found between Hungarian and Greek.

5.4 Discussion

To investigate the properties of Zipf's law cross-linguistically, Zipf's law was analyzed in fluent and non-fluent aphasic speech in three languages: English, Greek and Hungarian. For English, also a group of healthy control speech was available. First, all groups were compared to large, independent corpora of spoken language in the respective languages. Next, a within-language comparison was performed for the three groups of English speakers, and a between-language comparison was performed for the two aphasic groups in all three languages.

5.4.1 Baseline corpora vs. AphasiaBank samples

The speech samples that could be obtained from the AphasiaBank corpus were very small, containing only 200 words. Unfortunately, larger samples were not available, or would result in an even smaller number of speakers per group. In Chapter 3, it was argued that a minimum of 1500 tokens is required for calculations of Zipf's law. This would mean that the whole AphasiaBank corpus is unsuitable for any calculations concerning Zipf's law. This conclusion is unsatisfactory: cross-linguistic sources of aphasic speech are scarce and require an enormous amount of time and effort to acquire. It was thus attempted to find a method to nevertheless make the most of what is available, which is a highly valuable source of aphasic speech in multiple languages.

The way in which this was achieved was by analyzing baseline corpora. Zipf's law is language dependent, so different baseline corpora for each language were consulted. For each language, a corpus of spontaneous speech samples was sought, from which large numbers of 200-word samples were analyzed. The outcomes of these analyses formed a baseline, allowing us to inspect how much variation is to be expected for such small text samples, and also how the values that are found relate to the values for the language in general. The variation found in the AphasiaBank samples can thus be interpreted in light of this baseline.

The findings from the aphasic speakers were compared to the outcomes of the baseline corpora. Fit to Zipf-Mandelbrot's law in terms of R^2 was for none of the AphasiaBank groups lower than for the baseline corpora. In fact, for the English fluent aphasic speakers and for the Greek aphasic speakers[22], fit was better for the AphasiaBank samples than for the baseline samples. This means that the formula of Zipf-Mandelbrot's law allows for a more precise description of the aphasic speech samples than of the (healthy) baseline speech samples. The reason for this is unclear. In any case, the better fit (in combination with ZM-α values that fall within the normal range) allows us to safely conclude that speech from people with aphasia conforms to Zipf's law. This means that although the speech from fluent aphasic people and non-fluent aphasic people sounds markedly different from the speech of healthy speakers, this difference is not reflected in the fit of their frequency distributions to Zipf-Mandelbrot's law. These results confirm the findings in Chapter 4 and the earlier findings by Van Egmond et al. (2015) for Dutch non-fluent aphasic speech. More importantly, they provide evidence that this finding applies cross-linguistically and to fluent aphasic speech.

The conclusion that Zipf's law holds renders it safe to continue with an inspection of the parameter values. If variation within the large baseline corpora of healthy speech is so large that it incorporates the values found for aphasic speakers then it is questionable to what extent any differences between the two groups are meaningful. If, on the other hand, the values for the aphasic speakers fall outside of the normal range determined by the baseline corpus, then this is extra evidence that any difference is meaningful and real. The latter turns out to be the case. All groups of aphasic speakers, both fluent and non-fluent and in all three languages, differ significantly from the baseline corpora. In all cases, they display higher values of ZM-α irrespective of the type of aphasia. Strikingly, the English control speakers from the AphasiaBank corpus do not display significantly different ZM-α values.

From these analyses, it seems safe to conclude that Zipf-Mandelbrot's law does not apply any less to the AphasiaBank samples than to the corpora of the baseline samples. In addition, the variation found within the baseline corpora is not such that it incorporates the AphasiaBank samples. The ZM-α values found for the groups of

[22] And possibly the Hungarian fluent and non-fluent aphasic speakers: this effect was significant in the groupwise comparison (ANOVA) but failed to reach significance in the pairwise (post hoc) comparisons.

aphasic speakers are in fact in all cases higher than those of the baseline corpora. These differences can thus be considered to be due to the conditions of these speakers.

One more thing remains to be said with respect to the baseline corpora. Using this approach, it *is* in fact possible to use smaller-than-optimal text samples, as long as they can be interpreted in light of a larger corpus to correct for measurement errors due to the small samples. What you cannot say is that the values that are being found mean anything for the individual texts under consideration: the value of the parameter will likely grow when larger text sizes are considered (which is what was shown in Chapter 3). But the thing that *can* be done is using the parameter values to distinguish amongst groups. This is an important finding for clinical applications and studies of spontaneous speech, in which longer text samples are often unavailable or cumbersome to acquire and analyse.

5.4.2 Within-language Analysis

The finding that the parameter values of the AphasiaBank samples are not covered up by variation in the much larger baseline corpora made it possible to compare the AphasiaBank samples amongst each other. The first comparison in this respect was the one within the English language, since English is the only language for which the AphasiaBank corpus contained speech from healthy control speakers. Only for this language, a comparison between three groups was possible.

The statistical analysis of the ZM-α values showed that both fluent and non-fluent aphasic speech displayed higher values and thus a steeper slope compared to healthy control speech. These results show that although the speech from all three groups conforms to Zipf-Mandelbrot's law, the linguistic impairments of the two aphasic groups still affect the parameters of their word frequency distribution. The steeper slope of the non-fluent aphasic speech is in line with the van Egmond et al. (2015) findings for Dutch (see above, Section 5.1), thus validating the results found there. These results are stronger than those in the previous chapter, where no difference in slope was found between the groups as a whole. However, this seemed to be due to the mild nature of some of the aphasic speakers, since the more severely impaired individuals did seem to display steeper slope values (unfortunately, small sample sizes hindered statistical testing of this difference).

The results for the fluent aphasic speakers were not as expected. The healthy control speech was expected to have a steeper slope compared to the fluent aphasic speech, reflecting a more varied vocabulary for the fluent aphasic speakers due to neologisms and corrections. However, the results showed a significant effect in the opposite direction. This means that both groups of aphasic speakers performed the same way. As this is the first study investigating Zipf-Mandelbrot's law in fluent aphasic speech, further research is required to find out why this might be so. Zipf-Mandelbrot's law only concerns quantitative data. However, a qualitative investigation into the difference between fluent and non-fluent aphasic speech in light of these findings may be needed to explain them. Such a line of research should try to identify differences in the nature (i.e. neologisms and paraphasias versus existing words) and the linguistic properties (i.e. the linguistic category) of the items the two aphasic groups are producing. It might be the case, for example, that the small speech samples (i.e. 200 words) do not include as many paraphasias and neologisms for the fluent aphasic group and as many repetitions for the non-fluent aphasic group as needed to yield the predicted difference. Further research is imperative to shed light on these matters.

5.4.3 Between-languages Analysis

The second comparison within the AphasiaBank corpus was the one between fluent and non-fluent aphasic speakers in three different languages: English, Greek and Hungarian. The comparisons of ZM-α revealed a significant effect of language. Specifically, English has a steeper slope than Greek, and a trend for a steeper slope compared to Hungarian (p = 0,06), while Greek and Hungarian showed no significant difference between them.

Bentz et al. (2014), in a comparison between Old English and Modern English, showed that a language with poor morphology displays a steeper slope (i.e. higher alpha values) than a language with rich morphology. Thus, the morphologically poor Modern English has a steeper slope than the richer Old English. We suggest that the current results reflect this difference in the morphology of the three languages we studied. English is a morphologically poor language, especially when compared to Greek (and Hungarian). In Greek, nouns and adjectives are marked for number, gender and case, while verbs are marked for tense, person and number. However, in English, nouns and adjectives are only marked for number, and verbs are (only to some extent) marked for tense and person. This difference in grammatical encoding strategies is mirrored in the length difference of the tails of the frequency distributions, with English having a significantly steeper slope than Greek.

The reason why we did not find a statistically significant difference (although close at p = 0,06) between English and Hungarian is unclear. It seems to be the case that the difference is more pronounced for non-fluent aphasic speakers than for fluent aphasic speakers (the difference in slope was 0,076 for fluent aphasic speakers and 0,239 for non-fluent speakers), although no interaction effect with group was found.

The differences found in the between-languages analysis are in line with the slope differences found for the baseline corpora. Here, too, Greek and English have the largest difference between them: for Greek, a mean of ZM-α = 0,492 is found, while for English a mean of ZM-α = 0,724 is found. Hungarian is in the middle at ZM-α = 0,561. More research (preferably with larger groups) is necessary to see if it is the language morphology that causes the lack of difference, or if it has to do with the language impairments of the speakers. Greek and Hungarian, both morphologically rich languages, show no significant difference in ZM-α values between them.

Contrary to our predictions, but similar to the findings for the within-language analysis, the comparison of ZM-α between fluent and non-fluent aphasic speech revealed no significant difference between the two. As discussed before, further qualitative research is needed into the nature and the linguistic properties of the output of the two groups, to identify the exact differences and similarities between them.

5.5 Conclusion

The focus of the current chapter was the frequency distribution properties of aphasic speech, both fluent and non-fluent, in English, Greek and Hungarian. In the previous chapter and by Van Egmond et al. (2015), it was reported that Dutch non-fluent aphasic speech conforms to Zipf 's law. These were the first studies to provide this kind of evidence for any type of aphasic speech. The current study shows that it is not only non-fluent but also fluent aphasic speech that conforms to Zipf-Mandelbrot's law, and that this applies across languages with varying morphological complexities.

The variations in slope reported in this chapter between the non-fluent aphasic and healthy speech – with non-fluent aphasic speech having a steeper slope than healthy speech – are in line with the particular language deficits in this group. However, the finding that the fluent aphasic speech also had a steeper slope than the healthy speech was unexpected. A qualitative investigation into the difference between fluent and non-fluent aphasic speech in light of these findings is needed to explain them. The variations in slope between English and Greek appear to reflect the language-specific morphological properties of the two languages. This is supported by the trending difference between English and Hungarian. At the same time, we would also like to see whether the frequency distributions of languages of increasing morphological complexity show scalar differences in the parameters of Zipf-Mandelbrot's law, an interesting topic for further research.

Based on these findings, we can continue in two different directions, both equally important. On the one hand, we need to better understand what makes the speech of the aphasic population sound so distinctively different from healthy speech. It is remarkable that despite their impairments, the word frequency distributions of their speech continue to conform to Zipf's law. Any hypothesis for their difficulties should be compatible with this finding. On the other hand, these findings raise more questions regarding the lexicon and its properties on a more general, cognitive level. How is the lexical network organized, and how does this relate to the word frequency distributions of its output? These two lines of research complement each other, as each step forward in one will help to better understand the other.

To conclude, this is the first study to provide cross-linguistic evidence that aphasic speech, both fluent and non-fluent, although sounding markedly different from healthy speech, conforms to Zipf-Mandelbrot's law. These findings provide an extra step in revealing what exactly is impaired in aphasia, independently of its type. If the language system of an aphasic person is completely disrupted it would not adhere to fundamental aspects of natural speech, such as the Zipfian distribution of word-frequency. As was shown in the present study, this is not the case. This point will be discussed further in the next chapter, the General Discussion.

6 General discussion

6.1 Introduction

In Chapter 1, I formulated eight research questions. The first four related to Zipf's law in general, the other four were related to Zipf's law in aphasia. In the current chapter, I will give a short summary of the research reported on in the previous chapters and I will try to answer these research questions. This means that, where necessary, I will repeat results and conclusions from previous chapters to aid understanding, supplemented with new insights. Along the way, I will discuss some openings for further research. I will also reflect on the hopes I formulated in Chapter 1, to see if I got any closer to those ultimate goals.

6.2 Answers

With respect to Zipf's law in general

1. **What are the current hypotheses for Zipf's law, and which one seems to have most potential?**

A good explanation of Zipf's law explains why word frequencies follow the distribution that they do, both in shape and in slope, should be compatible with what we know about language production and should be falsifiable. Five branches of explanations for Zipf's law were discussed: the classic explanation of the Principle of Least Effort, explanations based on intermittent silence, the more modern approaches of preferential attachment and new optimization models building onto Zipf's Principle of Least Effort, and the varied group of explanations involving semantics.

The Principle of Least Effort was found to be too undefined to have full explanatory adequacy. It is intuitively appealing but lacks mathematical support. New optimization models fill this gap, mostly by providing models of semantic growth in which optimization processes account for a Zipfian output. However, these models seem to depend heavily on specific settings of the computational models involved. This means that slightly different settings result in different (possibly non-Zipfian) output. Intermittent silence models (such as Miller's (1957) theory about a monkey

Terminology on Zipf's law

The terminology used in this chapter is the same as elsewhere in this book, but is repeated here for convenience and in an attempt to limit confusion to a minimum (see also Chapter 2, section 2.2.1).
The term ‘Zipf's law’ is used as an overarching term for two formulas:

Zipf-Mandelbrot’s law:

9) $$f(w) = \frac{C}{(r(w) + \beta)^{\alpha}}$$

where the frequency *f* of word *w* is determined by its rank *r* when all words are ordered from most to least frequent, the parameters α and β, and a text size dependent constant *C*, and

Zipf’s β-law:

10) $$n_f = C \cdot f^{-\beta}$$

where the number of words *n* with frequency *f*, in other words, the size of the frequency class, is determined by parameter β and a text size dependent constant *C*.

Both Zipf-Mandelbrot’s law and Zipf’s β-law are formulations of the same phenomenon, and the unspecific term ‘Zipf's law’ has throughout literature been used to describe both. To avoid confusion, when either of the two laws is discussed specifically it will be called by its full name.
Another potential point of confusion is the parameter β, because both laws have one. Therefore, the β from Zipf-Mandelbrot’s law will be dubbed ZM-β, while the β from Zipf’s β-law will be dubbed Z-β.

hitting a typewriter at random) can be used as a null-hypothesis concerning Zipf's law in the absence of any other forces, but they are unrealistic as true models of language production: real texts do not consist of random strings of letters and conform to all sorts of syntactic requirements.
The varied branch of semantic models takes a very different approach to explaining Zipf's law. But except for Lestrade’s explanation regarding the combination of syntax and semantics, semantic models are not yet able to reproduce Zipf's law as it is found in human language. They can therefore not (yet) be relied on to explain the existence of Zipf's law. Lestrade’s model is the most promising semantic model. Unfortunately, Lestrade fails to provide any parameter values or measures for goodness of fit, which renders it difficult to properly evaluate this model.
Most promising and close to psychological (and possibly, neurological) reality is the group of explanations in terms of preferential attachment during network growth. These models show how Zipf's law can originate in a network as a result of the growth processes of this network. The relevant property of these growth processes is

preferential attachment. The structure of the networks that were modelled according to these processes is highly comparable to that of semantic networks built based on different sources of real-world semantic knowledge. In this way, these models are not only capable of explaining Zipf's law in natural language but can also explain similar phenomena in other fields in which similar frequency distributions are found: besides language, other phenomena can also be structured through growing networks (e.g. the World-Wide Web).

2. **How does Zipf's law behave for very small text sizes, and is there a lower limit in terms of number of tokens below which Zipf's law does not hold?**
3. **Are there any systematic differences in terms of Zipf's law between spoken and written texts?**

These two questions were addressed simultaneously in the first part of Chapter 3: the texts used to investigate the difference between written and spoken texts were also used to investigate the behaviour of Zipf's law in very small text samples.

The distinction between written and spoken language is not as clear-cut as it seems. Speeches are spoken, but often extensively edited and thought through before that time. Blogs, on the other hand, are written, but often not extensively edited which means that they are close to written down spoken language. Spoken or written language is a scale rather than a dichotomy. To reflect this fact, texts studied in Chapter 3 were selected from seven different categories that together form a continuum from spontaneously spoken to extensively thought through written text. The categories that were included are spontaneous face-to-face conversations, spontaneous commentaries on radio or television (which are unprepared but by trained speakers), more or less prepared radio and television discussions, and preaches and speeches in the group of spoken language; and blogs, news articles (written under time pressure so unlikely to be highly edited, but by trained writers) and literature in the group of written language.

Growth curves of each text were constructed for the different parameters of Zipf's law, ZM-α, ZM-β and Z-β (in other words, Zipf-Mandelbrots law and Zipf’s β-law, see the box at the beginning of this chapter or Chapter 2, Section 2.2.1). This was achieved by calculating the parameters for increasingly large parts of the text. It was attempted to fit a formula to these curves. There were three possible outcomes of this curve fitting procedure: a fitting formula, meaning that all growth curves for that parameter developed in an identical way; no fitting formula because of initial fluctuations for the smallest samples; or no fitting formula because values did not stabilize. All three options were found: for ZM-α I was able to find a well-fitting

formula, for Z-β it was found that initial fluctuations prevented a uniform description in the form of a formula and for ZM-β it was found that the parameter did not stabilize.
The growth curves for ZM-α were found to behave the same for all texts. In all cases, ZM-α quickly increased for very small samples, after which the increase levelled off and continued to grow at a slow but seemingly constant rate. The shape of the growth curve could accurately be described with a logarithmic function. Things look very different for the parameter ZM-β, for which no particular pattern could be discovered in the shape of the growth curves and values were found to cover a wide range from ZM-β = –0,8 to ZM-β = 10,5. The other formula for which growth curves were constructed was Zipf's β-law. A visual comparison showed rather similar growth curves: Z-β either decreased slightly when text size grows or stayed relatively stable. Nevertheless, no pattern could be discovered when comparing them statistically.

The reason for the large fluctuations of ZM-β probably lies in the nature of this parameter: it was introduced by Benoit Mandelbrot to obtain a better fit of Zipf's law for the highest rank numbers (Mandelbrot, 1954). ZM-β is calculated based on only a handful of ranks and thus data points, especially with small sized samples as used here. This renders the value of ZM-β prone to fluctuations, which explains the wide range of values and variation in growth curves found here.
These large fluctuations do not mean that the parameter is useless. On the contrary: including ZM-β in the formula of Zipf's law assures a more accurate calculation of ZM-α: thanks to ZM-β, ZM-α is less influenced by the deviation from a straight line in the first few ranks. What it does mean is that ZM-β is rather useless as a characterization of a text and should not be used as such. It is not possible to use ZM-β for comparisons amongst texts.

For Z-β, it is expected that a pattern can be found when larger text sizes are analysed. However, although it has to be concluded that no general pattern of growth could be discovered, it is still the case that values of Z-β stay relatively stable as text size increases. Parameter values do not largely fluctuate. This means that despite the absence of uniform growth curves, it can still be considered meaningful to compare the values of Z-β for texts with identical token counts of different sources.

The growth curves for ZM-α and for Z-β were used to determine the minimal workable text size, the text size at which the rate with which the value of the parameter changes stabilizes. It turns out that this happens around 1500 tokens. This means that parameter values calculated for texts of this size or larger can be used for comparisons between texts or groups of texts, as long as equal sample sizes are compared. This number lies at 1500 tokens, irrespective of the place of the text on the continuum from spoken to written language.

As already mentioned above, parameter values continue to grow when text size increases. This finding once again confirms that the parameters of Zipf's law are dependent on the size of the text. Zipf's law can thus only be compared between (groups of) texts when text sizes are equal.

Using equal sample sizes, written and spoken texts were compared. It was found that parameter values of ZM-α and Z-β reflected a larger vocabulary in written texts than in spoken texts, even if the spoken texts were prepared beforehand (e.g. sermons/speeches). ZM-α was higher for spoken texts than for written texts, Z-β was lower. These word frequency distributions suggest that the writers of these texts take into account that their text is going to be spoken aloud. A larger vocabulary is thus not a sign of the amount to which the text was prepared. Rather, it seems to be the case that a smaller vocabulary is necessary to keep the spoken text accessible. This makes sense, considering that the listener cannot pause, slow down or repeat part of the text as a reader can. This is (either consciously or unconsciously) taken into account when a spoken text is prepared.
ZM-β did not properly reflect the difference between written and spoken texts. Given the large fluctuations in its growth curves, this was expected.

Zipf's law thus does not behave radically different for written or spoken texts. But parameter values are not exactly the same either. ZM-α is generally higher for spoken texts; Z-β is lower. This finding confirms the correctness to treat ZM-α and Z-β as parameters rather than constants, an approach that is now usually (but still not always, e.g. Yang, 2013) the general practice.

Now, it is possible to answer questions 2 and 3. Question 2 concerned the presence of Zipf's law in very small text sizes. Fitting Zipf's law to very small text sizes means that the parameter values are highly influenced by every new token that is being introduced. A newly introduced type generates a new rank, which, when the total number of ranks is low, is of large influence on the total frequency distribution by lowering the slope. Repetition of a previously used token can – dependent on where in the distribution it is added – singlehandedly increase the slope. The more tokens already analysed, the smaller the influence of the next analysed token. This large influence of single tokens means that the parameter values that are being calculated for very small texts are not characteristic of the frequency distribution of the text, but rather are influenced by textual idiosyncrasies. Only after the initial 1500 words or so, this influence of individual tokens is decreased to such an extent that parameter values can be considered characteristic of the text as a whole. This does not mean that parameter values do not change anymore when text size is

increased beyond this point, but rather that it is now meaningful to compare the parameter values for texts from different sources. The number of 1500 tokens can thus be considered the lower limit beyond which the parameter of Zipf's law should not be considered characteristic of the text.
This is not the same as claiming that below 1500 tokens Zipf's law does not hold. It does: also for smaller text samples, it is found that ranked frequencies fall on a straight line. If Zipf's law holds for a larger text, then it also holds for any of its subparts, no matter how small. This becomes especially clear from Chapter 5, where text samples as small as 200 tokens were used. Even for these small samples, R^2-values for Zipf-Mandelbrots law (which are an indication of the fit of the data to the theoretical distribution of the law on a scale from 0 to 1) are at least $R^2 = 0{,}9$ or higher. The smallest texts that were analysed for the growth curves in Chapter 3 were 100 tokens. Even here, the lowest R^2-value that was found was $R^2 = 0{,}826$ and in most cases, it was well above $R^2 = 0{,}9$.

The problem is thus not if Zipf's law holds for a small text, but if the circumstances are such that it is possible to study Zipf's law in it. For Z-*β*, at some point, there are too many empty frequency classes for Zipf's law to be studied. For Zipf-Mandelbrots law this is not a problem. Nevertheless, for very small samples the problem is that any newly encountered word has a large influence on its frequency distribution as a whole. Somewhat larger text samples are required to overcome this issue. The values of Zipf's law when analysed in very small samples are not representative of the text as a whole, meaning that it does not allow for individual comparisons between texts. For these purposes, at least 1500 tokens are required.

Question 3 concerned the difference between written and spoken texts. The way in which ZM-*α* and Z-*β* developed when text size was increased did not differ for texts from different sources: the growth curves for all texts were highly comparable. This means that the parameters behave in the same way, irrespective of the source of the text. The exact values of the parameters, however, did depend on the source of the text. Generally speaking, ZM-*α* is lower and Z-*β* is higher for written texts than for spoken texts. This difference reflects the more varied choice of words in written text than in spoken text. Written or spoken is no dichotomy, but a continuum. This was reflected by the fact that texts close to the border between written and spoken (e.g. speeches, blogs) did not differ as much as texts on the edges of the continuum (e.g. spontaneous conversations, literature).

There is one more point that I would like to address here. Was G.K. Zipf wrong by stating that $\alpha = 1$? No, he was not wrong, but it is not the whole story.[23] In Chapter 3, it was found that the value of ZM-α (which is very close to the value of α in the traditional Zipf's law, see Chapter 2, Section 2.2.1 for a discussion of the formulas) grows logarithmically when text size is increased. The result of this is that the value of this parameter changes very little for a very wide range of text sizes. If the growth of ZM-α continues at exactly the same rate as given by the growth curves fitted in Chapter 3, then a value of which it can be said that it is approximately 1 can be fitted to text sizes up to 10.000 tokens (see Chapter 3, Section 3.2.3 for the calculations). This is where the idea of a constant comes from: there is none, but it seems like it because ZM-α values change little for such a wide range of number of tokens.

4. a. Is it possible to create a text for which Zipf's law does not hold?

The omnipresence of Zipf's law raises another question: does it *have* to be there, or can we create a text for which Zipf's law does *not* hold? There are many ways in which this question potentially could be answered, from which I chose one to pursue here. Deviations from Zipf's law are better detected when Zipf's β-law is considered than when Zipf-Mandelbrots law is considered, due to the absence of a ranking variable in this formula (see also Chapter 2, section 2.2.1). The ranking that is part of Zipf-Mandelbrots law forces the distribution to have a declining slope. Deviations can be detected when visually inspecting the distribution, but they do not necessarily show up in the values of the parameters or the fit of the distribution. This is different for Zipf's β-law. This formulation of the law is based on the size of the frequency classes, so in theory this distribution can take any shape. In practice, it follows a log-linearly declining distribution when Zipf's law applies. This typical distribution was disrupted here to obtain the non-Zipfian text: an artificial word frequency distribution was calculated in which the size of the word frequency classes followed a log-normal distribution instead. The text was then manually adapted to fit this new word frequency distribution. This was done for lexical words only: disrupting the distribution of grammatical words would disrupt syntax, thus making it impossible to know what you are measuring. The result of these manipulations was that the value of Z-β was about half that of the normal value for texts of this size (Z-β = 1,4 while Z-β = 2,2 for the original text). The biggest pointer to a disrupted distribution, however, was the fit to Zipf's β-law, Z-R^2 = 0,78, while usually values well above

[23] I would like to remind the reader that G.K. Zipf did his analyses, word counts and calculations without the help of computers.

0,9 are found. The values for Zipf-Mandelbrots law did not reveal the disruptions, although differences in the number of hapax legomena were visible from the plots of Zipf-Mandelbrots law. The calculation of Zipf-Mandelbrots law for lexical words only did show a difference: the value for ZM-β was much higher than usual (ZM-β = 204,4, while normally values above 10 are rare), and ZM-α = 2,5, more than six times as high as that of the control text (for which ZM-α = 0,4). These differences thus disappeared when lexical and grammatical words were combined in the analyses.

What this shows is that Zipf's β-law can be disrupted while Zipf-Mandelbrots law still displays rather normal parameter values. Looking at parameter values alone is thus not enough when this formulation of Zipf's law is concerned, and even the fit in terms of R^2 does not always reveal the disruption. The plot of Zipf-Mandelbrots law does show disruptions for the lower frequencies, but their influence on the values of the parameters is small. Zipf's β-law is much better at revealing the disruptions. The fit is reduced to almost 80% of its usual value, and the value of Z-β is halved. It is thus advised to use Zipf's β-law instead of Zipf-Mandelbrots law whenever the question arises if Zipf's law applies.

Combined, it seems safe to conclude that after the alterations, Zipf's law did not hold for lexical words. Whether or not it held for all words is debatable: the parameter values are within the normal range, yet the plots show some clear disruptions, especially in the number of hapax legomena. More drastic alterations thus seem to be necessary to remove Zipf's law from the text altogether. However, removing it at least from lexical words had never been attempted before. The fact that I managed to achieve this means that Zipf's law is not an absolute necessity when constructing a text. Readable texts *can* be constructed without Zipf's law, at least where content words are concerned. A different approach might allow for removal from function words as well, for instance constructing a new text from scratch instead of altering an existing text. This would be an interesting starting point for further research.

b. How do readers evaluate such a text?

Constructing a text for which Zipf's law does not hold is one thing, but reading such a text is another. Can readers detect the changes that were made? How do they evaluate the adapted text?

A comparison between an altered text and its original would not be fair, so I first constructed a control text. This text was formed by replacing words in the original text by their (near) synonyms, by changing some long sentences into short ones and vice versa, and by replacing non-stressed pronouns (‘ze’, ‘me’) by their stressed counterparts (‘zij’, ‘mij’). Parameter values of this text were close to the original ones, plots looked normal.

A questionnaire was designed to compare the texts. Three groups of participants were included, which each read and evaluated one of the three texts. The questionnaire addressed as many text characteristics as possible, since no expectations were present about the kind of differences readers would experience. Participants gave their first impression, they rated the text on a number of aspects, they rated the perceived variance in word choice, they judged the amount of perceived errors and they gave the text a general mark on a scale from 1 to 10.

People who read the non-Zipfian text did notice that the variation of words was smaller than that in normal texts. They did not particularly like the text, but it was not disliked more than the text that was modified in a different way. The one question in which the non-Zipfian text stood out was the question concerning how poetic the text was. Participants judged the text without Zipf's law to be significantly less poetic. The reason for this is unclear. However, poetry is generally known for its creative use of words. The text without Zipf's law was manipulated such that the same words were used more often. This means that word choice is more monotonous than usual. This unvaried choice of words is what likely caused this perceived difference between texts. Further research is needed to find out if this is indeed the case.

It seems to be the case that people (unsurprisingly) simply do not like modified texts. But the underlying reason for this is unclear: did participants dislike the text because it does not conform with every other text they ever read, or because it goes against their innate system? People always structure their speech such that Zipf’s law holds. In our lives, we have only heard texts for which Zipf’s law holds. It might be that this input taught us that that this is how one constructs a text. So far, Zipf’s law is found in every text in every language for which it has been tested.[24] However, if it were the result of the input, then one would expect that there would be at least one (possibly isolated) language for which Zipf’s law does not hold. Yet, the first language for which this is the case still has to be found. In addition, it might be expected that Zipf’s law does not always hold for the speech of very young children. Alternatively, Zipf’s law is hardwired from the start, because of organizational or

[24] The only exception being Modern Standardized Chinese, discussed in Chapter 2, Section 2.2.1. However, it seems to be that it is not Chinese perse that causes this, but the length of the texts in combination with the limited number of characters. Short texts or phrases probably do follow Zipf's law (see also footnote 14).

processing properties of the human brain. If so, then people did not like the manipulated text because it went against their innate organization. If this is so, then it is expected that even in child language Zipf's law always holds. Exceptions should not exist. Zipf's law in child language is studied very little. One study into this topic was performed by Baixeries, Elvevåg and Ferrer i Cancho (2013). They studied speech samples from children from ca. 18 months to 5 years of age with a focus on the period between 1-3 years, and from the adults they interacted with. They specifically looked at the development of the parameter α in the traditional Zipf's law. What they found is a decrease in α with time that is substantially stronger for the children than for the adults, suggesting that the children are not simply mirroring the behaviour of the adults with whom they are interacting. This supports the view that Zipf's law is hardwired. Unfortunately, no values for goodness of fit are given, thus rendering it impossible to evaluate if Zipf's law really holds in all cases.

More experiments with more different texts are needed to know exactly why people disliked the text.

With respect to Zipf's law and aphasia

5. a. Are there any systematic differences in terms of Zipf's law between people with and without non-fluent aphasia?

b. Is there a difference in the distributions of
- AoA and/or
- frequency in the Dutch language in general
in the speech of people with and without non-fluent aphasia?

c. In addition, does any of the measures above correlate with the severity of non-fluent aphasia, and do we see any developments when people recover?

d. Does Bentz and colleagues' NFD measure add anything to the picture as it follows from the questions above?

To answer these questions, I collected spontaneous speech from four people with mild forms of aphasia, three people with more severe forms of aphasia and seven matched controls. The control participants were recorded once, the aphasic participants were recorded at 2, 5 and 8 months post onset. All participants performed a picture description task, a movie description task and participated in a free conversation, amounting to approximately forty-five minutes of speech per person. Although still small, this corpus is the largest of its kind to my knowledge.

The differences that were found between groups were small. When mild and more severe aphasic speakers were combined, no differences were found with the control

group. Both for Zipf-Mandelbrots law and for Zipf's β-law parameter values were not significantly different, irrespective of the testing moment.
However, comparing the more and less severely affected participants with each other showed that the values of ZM-α were higher for the participants with more severe forms of aphasia at all testing moments, without any overlap between the two groups. This result was mirrored in the values that were found for Z-β, which were in all cases lower for the speakers with more severe forms of aphasia when compared to the participants with mild or no aphasia. The lack of overlap between groups is striking and suggests that there might indeed be a difference between more and less severely affected aphasic speakers (due to the small sample sizes, this difference could not statistically be compared). No such difference was found for fit or ZM-β, for which the split according to severity does not change the picture in any way. This is in line with the findings from Chapter 3, where ZM-β was found to strongly fluctuate for small texts due to its nature as a correction for the highest few ranks. No difference was found across testing sessions. Further research with larger groups is necessary to see if the difference between more and less severely affected aphasic speakers holds.

The steeper slope in the speech from the more severely affected participants could be a result of different strategies in dealing with the reduced resources due to their aphasia. Bastiaanse and Jonkers (1997) studied verb retrieval in naming and spontaneous speech in people with non-fluent forms of aphasia. They found that if people with aphasia show relatively normal verb inflection, a skill that is notoriously difficult for them, this is at the cost of diversity: they produce fewer different verbs. Those who ignore verb inflections show a (close to) normal diversity of verbs in spontaneous speech (Bastiaanse & Jonkers, 1997). Both strategies reduce the number of resources that is required: either the resources for retrieving different words or the resources for inflecting them. Both approaches also have the same effect on Zipf's law: the number of different types (and thus ranks) is reduced, resulting in a steeper slope.[25]

The finding that people with aphasia struggle significantly more than healthy control participants when the required amount of resources is high (also in non-language related tasks) is in line with work by Van Ewijk (2013). Van Ewijk presented both aphasic and healthy control participants with visual displays of circles. Participants were asked to indicate if any of those circles was smaller than the others. She found that both aphasic and control participants responded slower when more circles were

[25] This finding strengthens the idea of using Zipf's law as a classification tool for Zipf's law, as will be discussed in the answer to Question 8.

involved and when the size difference was smaller, but the difference in response times was much larger for the aphasic participants than for the control participants. The aphasic participants were thus much more affected by the increased complexity than the control participants. The difference in slope of Zipf's law between the more and less severely aphasic participants could be a result of the same effect. Both healthy and aphasic speakers will use difficult words less frequently, but this process is amplified in aphasia.

However, there is one effect that the reduced resources do not have: they do not distort the general shape of the frequency distribution. Zipf's law continues to hold. The downward curvature for high frequency words is not systematically different for the two groups (as reflected by ZM-β) and neither is the fit of the model (R^2). This suggests that the underlying system of word storage and retrieval functions normally. In other words: generally speaking, people with aphasia seem to store and retrieve words in the same way as people without aphasia.[26]

To answer question 5b, age of acquisition norms and values for frequency in Dutch (via SUBTLEX: Keuleers, Brysbaert & New, 2010) were collected for those words for which they were available. I would like to stress that I do not make any claims about the age at which these particular participants acquired the words they use, but rather about the age at which the words they use are usually acquired by average speakers of Dutch.[27]

Words used by people with aphasia were found to be on average acquired somewhat earlier in life. This difference was more pronounced for people with more severe forms of aphasia. This finding is in line with the finding that early acquired words are better retained by aphasic speakers (Bastiaanse, Wieling & Wolthuis, 2015). It fits in with the hypothesis that words that are acquired early are more deeply embedded in the lexical network (as discussed above under Question 1), thus rendering them more resilient to damage. No difference was found between testing sessions.

[26] This might not seem to be of much interest, but for people with aphasia, it is. It means that they did not lose part of their knowledge. For their self-esteem, this means everything (as one of them explained to me).

[27] Actually, the AoA values reflect estimates of AoA: A group of healthy speakers of Dutch was asked at what age they thought they acquired that word. These ratings are frequently validated and found to correlate highly with the actual age at which words are acquired by children learning the language (De Moor, Ghyselinck & Brysbaert, 2000; Gilhooly & Gilhooly, 1980).

The results for AoA are different from the results for SUBTLEX frequency. This suggests that they are two separate effects: if the AoA effect were a confound of the frequency effect (or the other way around) then no such differences would be expected. For SUBTLEX frequency, no difference between people with and without aphasia was found at any testing session. However, there was a difference between testing sessions within the group of people with aphasia: frequency values of the words used in session 1 were found to be significantly higher than those in session 2 and 3. This difference persists both for mild and more severe aphasic speakers. This difference might be the result of recovery, since high frequency words in the mental lexicon can be considered to be closer to the threshold of lexical activation. There is no inherent reason why stroke narratives (which was the main topic of conversation in session 1) would elicit more high frequency words than stories about family life (main topics of conversation in session 2 and 3).
In any way, it seems safe to conclude that people with aphasia do not perform differently from healthy speakers when general frequency in Dutch is concerned.

The finding that differences exist between groups for AoA but not for SUBTLEX frequency is in line with previous findings. Across studies, it is generally found that the effect of AoA is stronger than that of frequency (see Juhasz, 2005, for a review), and this difference persists for people with aphasia (Bastiaanse, Wieling & Wolthuis, 2016). This could be because AoA has a larger influence on the structure of the emerging lexicon than frequency. Learning a new word means establishing new connections in the lexical network, thereby altering the network structure itself. Frequency, by nature, can only affect the strength of these connections but does not change the structure. Frequency is also a factor that strongly fluctuates during one’s lifetime. For children, words like mummy or nappy are highly frequent, while for adults, they are usually not (until they have young children of their own). It is not yet clear how these fluctuations influence word storage and retrieval.

The answer to question 5c is simple: no, the measures applied here do not correlate with the severity of the aphasia. The only exception might be general frequency in Dutch (SUBTLEX frequency) of the used words, but this difference was in the opposite direction of what was expected. The question is what this means. Differences between people with and without aphasia included here were small in general, if any were found at all. So, it seems likely that any differences were simply too small to show up between testing sessions. If it does reflect any deeper property of aphasia then it might be the case that the problems faced by people with aphasia in terms of Zipf's law and AoA are not altered by recovery, which means that they are permanent.

Generally speaking, hypotheses for the impairments of aphasic speakers fall into two categories: a knowledge or competence approach, in which it is assumed that knowledge is permanently lost when someone acquires aphasia (e.g. Grodzinsky, 1995); and a processing or performance approach, in which it is assumed that all knowledge is retained but that processing limitations hinder the access of more demanding items (e.g. Avrutin, 2006). The processing approach has recently been favoured (e.g. Burkhardt et al., 2008). The supposed nature of the impairment determines the supposed character of people's recovery. If people with aphasia suffer from reduced resources then recovery either consists of regaining some of these resources or learning alternative, less demanding approaches to speech production, thus allowing them to re-access their (otherwise unchanged) knowledge. On the other hand, if their brain damage caused loss of knowledge then this is permanent. Recovery, in that case, consists of re-learning lost knowledge.
The lack of difference between testing sessions on all measures applied here (except SUBTLEX frequency) suggests that participants did not make progress on these matters. They did not use more words with a later AoA, and their active vocabulary did not increase (reflected in the unchanged parameter values of Zipf's law). This means that either they did not recover (which is unlikely in the post-acute phase, and not according to my impression of their speech when testing them, but the absence of diagnostics at the end of the testing period means this option cannot be ruled out), or the measures applied here tap into areas of knowledge that were not regained. The latter is more in line with a knowledge approach than with a processing approach. There is also an alternative option. One factor that was not included in this study was the time that participants needed to perform the tasks. The full testing session lasted about an hour but was not cut-off if it lasted longer. The required time was not noted. It could be that in earlier sessions they needed more time to perform the same tasks, resulting in the same performance on the measures applied here. The data that was collected consists of spontaneous speech without time pressure. This is very different from controlled experiments, which are usually applied to evaluate performance. It might be that this methodological difference is responsible for the lack of difference found in the current study. More research is imperative to disentangle these matters.

The calculation of the Normalised Frequency Difference (NFD) that was developed by Bentz and colleagues (Bentz et al., 2017) did not add anything to the picture, which answers question 5d. It was found that the difference within the control transcripts (between the first and second half of the texts) was larger than the difference between the first and second testing session of the aphasic participants and between the controls and third session of the aphasic participants. Differences between groups of speakers within a language are apparently too subtle to be captured by the NFD.

6. **Can any differences found between Dutch healthy and aphasic speakers be generalized to other languages?**
7. **Are there any systematic differences between fluent and non-fluent aphasic speakers?**

The study reported in Chapter 5 was designed to address question 6 and 7 simultaneously. In this chapter, Zipf's law was analysed in fluent and non-fluent aphasic speech in three languages: English, Greek and Hungarian. For English, also a group of healthy controls was available (source of all samples: AphasiaBank: MacWhinney, Fromm, Forbes, & Holland, 2011)

Unfortunately, however, the available sample sizes were very small, no larger than 200 tokens. The data that was used came from the AphasiaBank corpus, a corpus of cross-linguistic aphasic speech. It would be a waste of resources if this valuable source of data could not be used, since obtaining such a corpus requires an enormous amount of time and effort. This problem was solved by using baseline corpora, allowing for an interpretation of the results in light of the variation found within a much larger corpus. For each language, a corpus of 200-word speech samples from healthy speakers was analysed for Zipf's law. This analysis provided insight into the normal range of values for speech samples of this size. If the values for aphasic speakers would fall within this range then it would be impossible to claim that they perform any differently from healthy speakers. On the other hand, if their values would fall outside of it then this would be a clear signal that something is different about their speech. The variation found in the AphasiaBank samples can thus be interpreted in light of the variation within the baseline samples. Because of the small samples, only Zipf-Mandelbrots law was studied (the frequency class distribution needed for Zipf's β-law would contain too many empty classes, see also Chapter 3, Section 3.2.3).

In all cases, it was found that Zipf's law held. In fact, ZM-R^2 was in many cases even higher for the aphasic speakers than for the baseline samples. This means that the formula of Zipf-Mandelbrot's law allows for a more precise description of the aphasic speech samples than of the (healthy) baseline speech samples, which seems to suggest that a better-than-normal fit is in fact indicative of a deviation from normal speech.[28] In any case, it is safe to conclude that Zipf's law holds in all cases.

[28] A possible reason for this might be that aphasic speakers use fewer hapax legomena. Due to the logarithmic scaling, low frequency words form horizontal lines at the high-rank end of the distribution. The presence of more low frequency words means that more data points fall outside of the straight line given by the fitted formula of Zipf's law. Unfortunately, I realized this only after finalization of the research presented here. Time

All groups aphasic speakers, both fluent and non-fluent and in all three languages, differed significantly from the baseline corpora. In all cases, they displayed higher values of ZM-α despite the type of aphasia. The English control speakers from the AphasiaBank corpus did not display significantly different ZM-α values, so the difference is unlikely to be due to the different sources of the samples.

The finding that the parameter values of the AphasiaBank samples were not covered up by variation in the much larger baseline corpora made it possible to compare the AphasiaBank samples amongst each other. First, the English samples were compared amongst each other, since English is the only language for which a control group was available. It was found that both fluent and non-fluent aphasic speech displayed higher ZM-α values compared to healthy control speech. This steeper slope for non-fluent aphasic speech is in line with the findings in Chapter 4 for Dutch. The results for the fluent aphasic speakers, however, were opposite to what was predicted. The healthy control speech was expected to have a steeper slope compared to the fluent aphasic speech, but the results show a significant effect in the opposite direction. This means that both groups of aphasic speakers performed the same way. As this is the first study investigating Zipf-Mandelbrot’s law in fluent aphasic speech, further research and possibly qualitative analyses of the different speech samples are required to find out why this might be so. There could be (minimally) two reasons for the lack of difference: people with fluent and non-fluent aphasia have the same problems when spontaneous speech production is concerned, or they have very different problems that surface in the same way when Zipf's law is concerned. The first explanation is very unlikely. The distinction between fluent and non-fluent aphasia is the broadest possible distinction, and general consensus exists that these two groups are different (e.g. Goodglass, 1993; Axer et al, 2000). The second explanation is more plausible. As mentioned above, already within the group of non-fluent aphasic speakers it is possible that different coping strategies result in the same output when Zipf's law is concerned. The same holds true for the difference between fluent and non-fluent speakers. It was found that both groups use less varied vocabulary compared to healthy speakers. As mentioned in Chapter 5, variability in finite verbs is reduced in fluent aphasic speech. Reduced variability results in fewer different types and thus ranks, thereby creating a steeper slope. The number of paraphasias and neologisms produced by these speakers is apparently not large enough to compensate for this reduced variability. Further (qualitative) research is imperative to reveal the exact reasons for the lack of difference.

pressure prevented me from testing it after all. So, further research is needed to see if this hypothesis is confirmed.

Second, fluent and non-fluent aphasic speakers in three different languages (English, Greek and Hungarian) are compared. English was found to display a steeper slope than Greek, and a trend (p = 0,06) for a steeper slope compared to Hungarian, while Greek and Hungarian showed no significant difference between them. This finding is in line with previous research by Bentz et al. (2014), showing that languages with poor morphology display a steeper slope. This finding thus extends to aphasic speech: English is generally considered to be morphologically poor, while Greek and Hungarian are considered to be morphologically rich languages.

Similar to the findings for the within-language analysis, no differences were found between the groups of aphasic speakers in the three languages. As discussed above, the reason for this remains a topic for further research.

8. Are there any possible clinical implications or possibilities for people with aphasia or other impairments that follow from the study of Zipf's law?

The output of a healthy language system follows Zipf’s law. Much less is known about the output of a damaged system, but up to now no examples have been found for which Zipf’s law does not hold. Nevertheless, clear differences are found when impaired populations are compared to control groups, especially in the slope of the distribution: schizophrenic patients with disconnected speech display a more gradual slope; schizophrenic patients with a topic of obsession display a curve instead of a straight line, and children with Down syndrome display a steeper slope (Piotrovskii, Pashkovskii & Piotrovskii, 1994; Piotrowski and Spivak, 2007). The current study showed that people with aphasia – irrespective of the type – display a steeper slope. One thing that becomes very clear from this is that – no matter how different their speech sounds – the basic properties of Zipf’s law remain intact. This suggests that the language faculty itself basically functions unimpaired, albeit under restrictions.

It remains to be seen what happens to Zipf’s law in populations that do suffer from an impairment to the core of the language structure by disrupting for instance basic linguistic structure or selection and (short-term) storage of lexical items. One disorder I would be particularly interested in is semantic dementia. This progressive neurodegenerative disorder is a subtype of frontotemporal dementia, also known as semantic variant primary progressive aphasia (svPPA). People with semantic dementia suffer from loss of semantic knowledge, resulting in word finding difficulties, difficulties in understanding words they previously knew, talking vaguely and not understanding people around them. Problems with the recognition of objects or faces and impairments of day-to-day events usually do not occur until in a later stage (FTD talk, Factsheet 5). The difference between semantic dementia and aphasia is that people with aphasia are thought to have retained their knowledge but suffer from language problems due to reduced resources, whereas people with semantic dementia do not suffer from reduced resources but rather from actual loss of knowledge. I hypothesize that this may actually – at least when the disorder has reached an advanced enough stage – disrupt the Zipfian distribution of word frequencies in their speech. If this is so then this would be strong evidence that Zipf’s law actually originates within the language system, as a result of the organization of the mental lexicon.

Another important and for speech therapy potentially relevant conclusion is that the patients that were included in this study were being treated by different speech therapists from different rehabilitation centres. In addition, the nature of their aphasia was for all slightly different, since after all aphasia is a very diverse condition. As a result, they all received different kinds of therapy. Nevertheless, they performed rather uniformly on the measures performed here. There were no individual aphasic speakers with aberrant parameter values. This study looked at changes in the aphasic lexicon during recovery, not the way in which these changes came to be. That means that it was not a requirement of this study to control for the kind of therapy that patients received. Therapy might be of some influence on the results but is not a threat to it: it is the reality of how actual patients recover. This study of Zipf’s law suggests that these individual differences do not influence the frequency distributions in patients’ speech.

From a more practical point of view, the fact that Zipf’s law can also be studied in very small samples means that Zipf’s law could potentially be used as an interesting addition to the classification of patients. In Chapter 4 it was found that the slope of the more severely aphasic speakers was steeper than that of the less severely aphasic speakers. This means that Zipf’s law has potential to be used as a theory free classification method. A steeper slope then indicates larger language problems. This method would allow for classification free of influence of factors such as poor pronunciation or long pauses.

Transcription and analysis of spontaneous speech is already frequently part of the classification of patients, for instance in methods such as the ASTA (Analyse voor Spontane Taal bij Afasie [Analysis for Spontaneous Language in Aphasia], Boxum, Van der Scheer & Zwaga, 2010). This method is designed for quantitative analysis of spontaneous speech in people with aphasia. Participants are interviewed according to a fixed protocol, resulting in transcripts of 300 words. These transcripts are then scored on a number of phonematic, morphosyntactic and lexical-semantic measures. With the right software, these transcripts can easily and without extra effort be analysed for Zipf’s law. A large number of samples should first serve to calibrate the method, to see how severity and slope correlate exactly. The fixed protocol of the ASTA should help minimize variation. Future research should reveal if this approach is feasible, and how much it adds to existing methods.

6.3 Hopes and dreams

In the introduction of this dissertation, I presented two things I hoped for. Did I get any closer to those ultimate goals?

My first hope was that someday, we will be able to pinpoint exactly where Zipf's law comes from. I reviewed a large body of existing literature to see how others explained it. From this, I concluded that it is most likely that it is the network in which the mental lexicon is structured that is responsible for the omnipresent nature of Zipf's law in language. I studied Zipf's law in small samples and in spoken language, which were prerequisites to be able to continue with the study of Zipf's law in impaired populations. I then continued with a study of Zipf's law in aphasia, one of the most well-known language disorders. The fact that Zipf's law holds in speech from people with aphasia no less than in speech from healthy speakers strengthens the idea that Zipf's law originates in the organization of the mental lexicon, because people with aphasia are thought to suffer from language problems due to reduced resources, not due to reduced knowledge. But more evidence is needed to know if this is true. This evidence should especially come from cases where the language system itself is impaired. If the organization of the lexical network is indeed the reason for Zipf's law then a disrupted lexical network should result in speech for which Zipf's law does not hold. Disrupted networks are hard to find, but one potential population would consist of people with semantic dementia in a more advanced stage. A study of this population would be the perfect continuation of the work presented here.

My second hope was that my research would someday be able to serve as a tool to gain more insight into the workings of language in a more general sense, and, more specifically, somehow help people with a language disorder. More work is needed before these goals are fully achieved, but I believe I made some steps in the right direction. I provided a little more insight into aphasia, mostly by showing that it actually is not too different from healthy speech. This might not sound like much, but to people with aphasia, it means the world: it shows that despite everything, they did not fundamentally change, and it helps them to keep hope for recovery because their knowledge is intact. This can be an important mental boost for them, as one of them explained to me.

My research made it plausible that Zipf’s law originates in the mental lexicon; other research is imperative to either verify or falsify my hypotheses. I expect the most of the study of people with semantic dementia, as explained above. If we can someday prove that Zipf’s law originates in the mental lexicon then this gives invaluable insight into the workings of that same mental lexicon. It would mean, conversely, that the mental lexicon in aphasia is (at least broadly) unharmed. This would be an important insight for the development of speech therapy for these patients: it would show that therapy should, where possible, focus even more on ways to boost the availability of resources, rather than re-learning hard-to-access lexical items.

In my answer to research question 8, I presented a way in which Zipf’s law could potentially be used to classify patients by analysing the output of the ASTA protocol for Zipf’s law. The method that I presented is within reach. Hopefully, someday, it will be possible to implement Zipf’s law in this way.

PS. Is Zipf’s law everywhere?

There is one more point that I would like to address. Throughout the literature, it has often been claimed that Zipf’s law is practically everywhere (e.g. Baek, Bernhardsson & Minnhagen, 2011; Corominas-Murtra & Solé, 2010; Dahui, Menghui & Zengru, 2005), in areas as diverse as city sizes (e.g. Zipf, 1949; Krugman, 1996), links on the World Wide Web and other features of the Internet (e.g. Adamic & Huberman, 2002), the income distribution of companies (Okuyamaa, Takayasub & Takayasuc, 1999) and gene expression on various organisms and tissues (Furusawa & Kaneko, 2003). Some claim, therefore, that Zipf’s law is uninteresting (e.g. Yang, 2013). But what they mean is not that Zipf’s law is everywhere, but that *power law behaviour* is everywhere. Zipf’s law has been used as an overarching term for power law behaviour in many different fields. Often, a slope of -1 happens to be found, comparable to that of Zipf’s law. However, this is *not* the same as Zipf’s law being everywhere. Zipf’s law as found in texts has a very particular shape: parameters are within a certain range (not too far away from $\alpha \approx 1$ and Z-$\beta \approx 2$), a downward curvature is found for the first few ranks, and there is a large percentage of hapax legomena. In this particular form, Zipf’s law is unique to language: especially this typical downward curvature is often missing from other areas for which it has been claimed that Zipf’s law holds. The fact that similar power laws are found in different areas does not mean that the reason for this behaviour is the same in all cases. Overlap in terminology does not mean that the processes at work are identical. It looks comparable, but it does not have to *be* the same.

Appendices

A Texts used in Chapter 3, part II

A1 Original text

1.

Het zal me benieuwen waar ze nu weer mee komt. Het verhaal van de vorige keer was zo lachwekkend dat ik er pijn van in mijn zij had. Ineens had mevrouw RSI-verschijnselen in haar knieën waardoor licht administratief werk onmogelijk zou zijn. Ze had gelijk gekregen van haar dokter en ik moest ook overstag. Dat was vorige week. Over een kwartier zal ze zich bij me melden, met een gedetailleerd rapport van de dokter. Ik wil wel eens zien met wat voor diagnose die kwakzalver me denkt in te pakken.

Mijn kamerdeur staat tijdelijk open. Om deze werkdag door te kunnen komen moest ik de boel flink luchten, want het rook bijzonder onfris toen ik vanmorgen binnenstapte. Alsof de schoonmaakploeg gisteravond op mijn kamer kebab had klaargemaakt. Het zou me trouwens niks verbazen als dat regelmatig gebeurt. Ik heb eens een paar koffiekringen in de vensterbank gemaakt om te kijken of die lui wel werken voor hun geld. Het was schokkend. Ik heb zes werkdagen tegen een vieze vensterbank moeten aankijken. Uiteindelijk heb ik het zelf weggeveegd met een doekje uit de keuken. Ongehoord. Wat is er mis met werkloosheid als je toch niet wilt werken voor je geld?

Rumoer op de gang. Een verzetje. Misschien is er iemand van het toilet gekomen die zich het Fresh Prince-grapje moet laten welgevallen, maar zo klinkt het niet. Dit heeft meer iets van een dialoog. Ik kijk om de hoek en zie Tineke opdoemen, vijf minuten eerder dan afgesproken. Ze is halverwege de gang staande gehouden door een collega, die vraagt of ze snel weer terugkomt. Ze weet het nog niet, maar ze heeft goede hoop, zegt ze. Daarna barst ze in een soort lachbui uit, die kenmerkend is voor haar genre.

Communicatief niet vaardig houden zij zich amper staande in een groep. Ze concentreren zich op een aandoenlijke manier op hun gesprekspartner en kunnen een antwoord vanuit hun tenen geven, waarna ze de opgebouwde spanning kwijtraken door hard te lachen en te hikken.

Binnen het Arbo-wezen wordt er altijd op gehamerd dat soort vrouwen kort te houden, want vanuit hun beperkte inlevingsvermogen zijn ze vatbaar voor waanideeën en dat kost de werkgever alleen maar geld.

Zul je altijd zien: zit zo'n mens in haar periode naar een infodocu over spierziektes te kijken en hup, ze voelt wat in haar nek.

En dan wil ze natuurlijk liever het zekere voor het onzekere nemen. Dat moet de Arbo-man dan maar snappen. Niet dat ze echt wat aan haar zogenaamde aandoening doet, want ze zit wél gewoon om de avond met zes kaarten bij de bingo. Nee, het eerste wat er bij dat soort mensen aan moet geloven is hun baan.

De inhoud van je loonzakje loopt toch geen gevaar. Dat is zo'n fijne verworvenheid van onze beschaving en de reden voor bedrijven om hun productie over te hevelen naar landen waar dingen in alle redelijkheid gebeuren, zonder uitputtende regelgeving die werknemers lui maakt.

Tineke staat nog steeds op de gang. De collega houdt haar aan de praat en zij komt op adem. Tineke heeft tamelijk forse tieten en de bewuste collega is dat nog niet vergeten. Puur en alleen voor zijn gemoedsrust zou ik Tineke snel weer op de werkvloer willen terugzien, al zou ik haar het liefst inruilen voor een Aziatisch type.

Die zeuren niet, zijn nooit ziek en ze weten volgens mij ook niet wat een Arbo-medewerker voor ze kan betekenen. Des te beter, want ik heb het druk genoeg.

Zat van het wachten draai ik me de gang op en roep Tineke tot de orde. Ik heb niet de héle dag de tijd, zeg.

Als mevrouw Burgers mijn kamer binnenkomt, heeft haar vrolijkheid van zonet plaatsgemaakt voor een imitatie van zakelijkheid. Ze heeft een gekleurde snelhechter in haar hand. In de gauwigheid zie ik grote rode en zwarte letters die er met viltstift zijn opgeschreven. Ik hoop dat ze een opstel voor me heeft, maar het blijkt haar revalidatierapport van de dokter te zijn.

'Leuk je weer te zien,' lieg ik haar tegemoet. 'Pak een stoel.'

'Dank je,' zegt ze formeel.

Als ze gaat zitten veer ik kort even op achter mijn bureau. 'Zo,' zeg ik opgewekt, 'dat ziet er bemoedigend uit. Je steunde bij het zitten gaan vol op je polsen.'

'Nee hoor, dat kan helemaal niet,' zegt ze als door een wesp gestoken.

'O ja zeker wel. Ik zag het je net zelf doen.'

'Uitgesloten.'

'Weet je wat ik denk? Ik denk dat je onderbewustzijn allang is gewend aan je voorwendselen.' Ze kijkt me even aan en steekt de snelhechter naar mij uit. 'Dit is het re...'

'Revalidatierapport ja, dat staat er in koeienletters op. Geef hier, ik ben reuze benieuwd.'

'Trouwens,' zegt ze koel, 'ik wend me eigen niets voor.'

'Staat dat hier ook in?' Voor ze een antwoord kan geven, heb ik haar al tot kalmte gemaand, want met dat getetter kan ik me niet concentreren op complexe medische materies, zoals dit rapport. Na een halve tel snap ik al helemaal niets meer van de opbouw van het betoog. Het is een rapport van niks, dit lijkt nog het meest op een hobbyroman van een plattelander. 'Die dokter van jou, hè, wat is dat eigenlijk voor man?'

‘Dokter Verspochten is een vrouw,’ bitst Tineke, ‘Els.’

‘Aaah,’ zeg ik op een jazzy toon, om vervolgens weer net te doen alsof ik aan het lezen ben. Precies wat ik dacht. Een vrouw. Zo’n mens dat patiëntes ronselt bij de batikclub en bij de vrouwen-gym. ‘Meid, het lijkt wel of je met je rug trekt. Joh, weet je wat? Kom morgenmiddag even bij me langs, kijk ik er even naar.’ Dat haar hobbyistische vriendinnenmanie uiteindelijk door de belastingbetaler moet worden opgehoest, interesseert Els natuurlijk geen dikke reet.

A2 Text without Zipf’s law in content words

1.

Ik ben benieuwd met welk goed verhaal mevrouw nu weer langs zal komen. Het verhaal van vorige week had me zo aan het lachen gemaakt dat ik er pijn van in mijn zij had. Toen had ze ineens pijn in haar knieën waardoor licht administratief werk onmogelijk zou zijn. Ze had geregeld dat haar dokter haar gelijk gaf en nu moest ik haar ook geloven. Dat was vorige week. Over een kwartier zal ze hier terug zijn, met een rapport van de dokter. Ik wil wel eens zien met welk verhaal die meneer me wil inpakken.

Mijn deur is open. Om deze dag beter te kunnen doorkomen moest ik mijn kamer goed luchten, want het rook bijzonder vies toen ik net was binnengestapt. Alsof de schoonmakers de vorige dag in mijn kamer kebab hadden gemaakt. Het zou me trouwens helemaal niks meer verbazen als ze dat zouden doen. Ik heb eens een paar koffiekringen achter op mijn bureau gemaakt om te kijken of de schoonmakers wel willen werken voor hun geld. Het was schokkend, echt schokkend. Ik had precies zes hele werkdagen een vies bureau. Uiteindelijk moest ik de gemaakte koffiekringen zelf van het bureau vegen. Waarom geen werkloosheid als ze toch niet willen werken voor hun geld?

Getetter op de gang. Is er een collega van het toilet gekomen waarna die zich het slechte Fresh Prince-grapje moet laten welgevallen? Dat lijkt het niet. Dit heeft meer iets van een dialoog. Ik kijk door de open deur en zie dat Tineke er nu al staat, een kwartier eerder dan afgesproken. Ze staat in de gang met een collega, die vraagt of hij haar snel zal terugzien. Ze weet het nog niet, maar ze heeft goede hoop, zegt ze. Dan barst ze in een soort lachen uit, dat kenmerkend is voor haar soort mensen.

Zij weten het nauwelijks uit te houden in een groep. Ze zijn gewend zich op een aandoenlijke manier op de dialoog te concentreren en een antwoord te geven vanuit hun tenen, waarna ze de opgebouwde spanning verliezen door schokkend te lachen.

In mijn werk is afgesproken dat soort mensen kort te houden, want vanuit hun slechte inlevingsvermogen zijn ze vatbaar voor waanideeën waardoor hun werkgevers alleen maar geld verliezen.

Zul je altijd zien: zit zo'n vrouw eens in een boek over spierziektes te lezen en hup, ze heeft ineens wat in haar knieën.

En dan wil ze natuurlijk liever haar gezondheid veilig stellen. Of ik dat niet snap. Niet dat ze trouwens echt iets aan haar slechte gezondheid doet, want ze zit wél steeds om de dag bij de bingo. Nee, het eerste wat er bij dat soort mensen aan moet geloven is hun werk.

Hun geld is toch veilig. Dat is kenmerkend voor ons land en een goede reden voor werkgevers om hun productie uiteindelijk over te hevelen naar een land waar de zaken beter gebeuren.

Tineke staat al die tijd nog steeds op de gang. De collega houdt haar aan de praat: Tineke heeft bijzonder grote tieten en de bewuste meneer weet dat natuurlijk ook nog. Alleen al voor zijn gezondheid zou ik Tineke nu liever snel weer hier willen terugzien, alleen zou ik haar liever vervangen door een Aziatische. Die zullen nauwelijks moeilijk doen, zullen er altijd zijn en ze weten denk ik ook niet wat ik voor ze kan doen. Des te beter, want ik heb werk zat.

Zat van dat getetter draai ik me de gang op en roep Tineke. Ik heb niet de héle dag de tijd, zeg ik.

De collega kijkt me kort aan en gaat snel terug naar zijn kamer. Als mevrouw mijn kamer is binnengestapt, heeft ze haar lachen van net ineens vervangen door gemaakte zakelijkheid. Ze heeft iets vast, en ik zie de grote rode letters die er zijn opgeschreven. Ik heb even de hoop dat ze een goed boek voor me heeft, maar het blijkt haar revalidatierapport te zijn.

'Goed je weer te zien,' zeg ik. 'Ga zitten.'

'Dank je,' zegt ze, nog altijd met die zakelijkheid.

Als ze gaat zitten veer ik kort op achter mijn bureau. 'Zo,' zeg ik met gemaakte vrolijkheid, 'dat is bemoedigend. Je steunde bij het zitten gaan goed op je knieën.'

'Nee hoor, onmogelijk,' zegt ze kortaf.

'O ja echt wel. Ik zie het je net zelf doen.'

'Onmogelijk,' zegt mevrouw net te snel.

'Weet je wat ik denk? Ik denk dat je al gewend bent aan je voorwendselen.' Ze kijkt me even aan en houdt het revalidatierapport naar mij uit. 'Dit is het re...'

'Ja dank je. Revalidatierapport ja, dat staat er in bijzonder grote, rode letters op. Kom hier, ik ben benieuwd.'

'Trouwens,' zegt ze kortaf, 'ik ben niet gewend aan voorwendselen.'

'Staat dat hier ook in?' Voor ze een antwoord kan geven, heb ik haar al tot kalmte gemaand, want met dat getetter kan ik me onmogelijk concentreren op moeilijk opgebouwde zaken, zoals dit rapport. Al snel snap ik helemaal niets meer van het rapport. Het is een rapport van niks, dit lijkt eerder op een eerste boek van een plattelander. 'Die dokter van jou, hè, wat is dat voor een meneer?'

'Mijn dokter is een mevrouw,' bitst ze, 'Els.'

'Aaah…,' zeg ik, 'Els,' om dan weer net te doen alsof ik aan het lezen ben. Precies wat ik dacht. Een vrouw. Vast zo'n goede vrouw die patiëntes ronselt bij de bingo en bij de vrouwen-gym. 'Joh, wat heb je met je knieën? Weet je wat? Kom van de week even bij me langs, dan kijk ik even.' Dat het geld voor de bewuste patiëntes uiteindelijk van de belastingbetaler moet komen interesseert Els natuurlijk helemaal niks.

A3 Control text

1.

Het zal mij benieuwen waar zij nu weer mee komt, want de vertelling van de vorige keer was zo komisch dat ik er pijn van in mijn zij had. Ineens had madam RSI-verschijnselen in haar knieën. Licht administratieve werkzaamheden zouden daarom onmogelijk zijn. Zij had gelijk gekregen van haar huisarts en ik moest ook overstag. Dat was vorige week. Over vijftien minuten zal zij zich bij mij melden, met een precies verslag van de huisarts. Ik wil wel eens kijken met wat voor uiteenzetting die kwakzalver mij denkt te lijmen.

Mijn kamerdeur staat kort open. Om deze werkdag te overleven moest ik het zooitje flink ventileren. De geur was bijzonder goor toen ik vanmorgen binnenstapte. Alsof de huishoudelijke dienst gisteravond op mijn kamer kebab had klaargemaakt. Het zou mij overigens niets verbazen als dat regelmatig plaatsvindt. Ik heb eens enkele koffiekringen in de vensterbank gemaakt om te kijken of die lui wel arbeid verrichten voor hun loon. Het was shockerend, want ik heb zes werkdagen tegen een vieze vensterbank moeten aankijken. Uiteindelijk heb ik het zelf weggepoetst met een lapje uit de keuken. Ongehoord. Wat is er verkeerd aan werkloosheid als je toch niet wilt werken voor je loon?

Lawaai op de gang: wat afleiding. Misschien is er een persoon van de wc gekomen die zich het Fresh Prince-geintje moet laten welgevallen. Maar zo klinkt het niet, want dit heeft meer iets van een gesprek. Ik kijk om de hoek. Ik zie Tineke verschijnen, vijf minuten eerder dan overeengekomen. Zij is halverwege de gang tot staan gebracht door een collega. Hij vraagt of zij snel weer terugkeert. Zij weet het nog niet, maar heeft goede moed, zegt zij. Daarna barst zij in een soort lachstuip uit, die kenmerkend is voor haar soort.

Communicatief niet bedreven houden zij zich nauwelijks staande in een gezelschap. Zij focussen zich op een schattige wijze op hun conversatiepartner en kunnen een reactie vanuit hun tenen geven, waarna zij de opgebouwde nervositeit verliezen door hard te grinniken en te hikken.

Binnen het Arbo-wezen wordt er altijd op gehamerd dergelijke individuen kort te houden, want vanuit hun gelimiteerde empathie zijn zij bevattelijk voor hersenspinsels en dat kost de werkgever uitsluitend centen.

Zul je altijd zien: zit zo’n vrouwtje tijdens haar menstruatie naar een infodocu over spierziektes te kijken en hup, zij voelt wat in haar nek.

En dan wil zij uiteraard liever het zekere voor het onzekere nemen, wat de Arbo-medewerker dan maar moet begrijpen. Niet dat zij daadwerkelijk wat aan haar ingebeelde ziekte doet, want zij zit wél gewoon om de avond met zes kaarten bij de bingo. Nee, het eerste wat er bij dat soort figuren aan moet geloven is hun baan, want de hoogte van het inkomen loopt toch geen risico.

Dat is zo’n fijne aanwinst van onze civilisatie en de reden voor corporaties om hun productie te verhuizen naar naties waar dingen in alle billijkheid plaatsvinden, zonder vermoeiende wetgeving die loontrekkers lui maakt.

Tineke bevindt zich nog steeds op de gang. De collega houdt haar aan de praat. Zij komt op adem. Tineke heeft tamelijk prominente borsten en de collega in kwestie is dat nog niet vergeten. Uitsluitend voor zijn zielenrust zou ik Tineke snel weer bij ons bedrijf willen weerzien. Al zou ik haar bij voorkeur inwisselen voor een Aziatisch type.

Die emmeren niet, zijn nooit ziek en zij weten volgens mij ook niet wat een Arbo-medewerker voor ze kan doen. Des te beter, want ik heb werk genoeg.

Moe van het oponthoud keer ik mij de gang op en roep Tineke tot de orde. Ik heb niet de héle ochtend de tijd, zeg.

Als mevrouw Burgers mijn kantoor binnenstapt, heeft haar opgewektheid van zonet plaatsgemaakt voor een nagemaakte professionaliteit. Zij heeft een gekleurde snelhechter vast, en in de gauwigheid zie ik grote rode en zwarte letters die er met viltstift zijn opgekrabbeld. Ik hoop dat zij een roman voor mij heeft, maar het blijkt haar revalidatierapport van de huisarts te zijn.

Aangenaam jou weer te zien,' jok ik haar tegemoet. Neem een stoel.'

Dank je,' zegt zij vormelijk.

Als zij gaat zitten richt ik me kort even op achter mijn schrijftafel. Zo,' zeg ik energiek, dat ziet er opwekkend uit. Jij rustte bij het zitten gaan vol op je polsen.'

Nee hoor, dat kan helemaal niet,' zegt zij als door een insect gestoken.

O ja zeker wel, want ik zag het jou net zelf doen.'

'Uitgesloten.'

Weet jij wat ik vermoed? Ik vermoed dat jouw onderbewustzijn allang gewoon is aan jouw voorwendselen.' Zij kijkt mij even aan. Zij steekt de snelhechter naar mij uit. Dit is het re...'

Revalidatierapport ja, dat staat er in gigantische letters op. Geef hier, ik ben enorm nieuwsgierig.'

Trouwens,' zegt zij beheerst, ik wend me eigen niets voor.'

Staat dat hier ook in?' Eer zij een antwoord kan geven, heb ik haar al tot kalmte gemaand. Met dat geklets kan ik mij niet focussen op ingewikkelde medische zaken, zoals dit rapport. Na een halve seconde snap ik al helemaal niets meer van de structuur van de uiteenzetting. Het is een rapport van niks, dit lijkt nog het meest op een hobbyroman van een boer. Die huisarts van jou, hè, wat is dat eigenlijk voor vent?'

Dokter Verspochten is een vrouw,’ snauwt Tineke, Els.’

Aaah,’ zeg ik op een jazzy toon, om daarna weer te imiteren dat ik aan het lezen ben. Een vrouw, precies wat ik dacht. Zo’n persoon die patiëntes aantrekt bij de batikclub en bij de vrouwen-gym. Lieverd, het lijkt wel of jij met je rug trekt. Joh, weet je wat? Kom morgenmiddag even bij mij aan, kijk ik er even naar.’ Dat haar hobbyistische vriendinnenmanie op den duur door de gewone burger moet worden betaald, boeit Els uiteraard helemaal niets.

B Questionnaire

Fijn dat je mee wilt werken aan dit onderzoek!

Deze tekst is door een beginnende vertaler vertaald uit het Engels. Wij willen graag weten hoe jij deze tekst beoordeelt.

Lees eerst op je gemak de tekst. Je hoeft de tekst niet uitgebreid te bestuderen, maar lees hem wel helemaal. Beoordeel daarna de tekst. Hiervoor hebben we een vragenlijst opgesteld, bestaande uit vier open vragen en 49 meerkeuzevragen. Geef een antwoord op alle vragen, in de volgorde waarin ze gegeven zijn. Denk niet te lang over je antwoorden na maar volg je eerste ingeving. Je mag de tekst er tijdens het geven van de antwoorden bij houden.

Je antwoorden zullen anoniem verwerkt worden.

Hartelijk dank namens de onderzoeker,
Marjolein van Egmond
M.vanEgmond1@uu.nl

Wil je op de hoogte gebracht worden van de uitkomsten van dit onderzoek? Vul dan hieronder je emailadres in. Je emailadres zal uitsluitend gebruikt worden voor dit doel en zal niet aan derden doorgegeven worden. Je emailadres zal apart van je tekstbeoordeling bewaard worden.

emailadres: ……………………………………………………………………………………………

Vragenlijst tekstbeoordeling

- Leeftijd: ………………………………
- Geslacht: m / v
- Moedertaal: NL Anders, nl ……………………………………………………

A.

1. Wat is je eerste indruk van deze tekst? Richt je in je antwoord vooral op wat je is opgevallen wat betreft aspecten zoals stijl, vorm, woordkeuze en taalgebruik, niet op de inhoud.

B.

Geef aan in hoeverre je het eens bent met de volgende stellingen. Doe dit op een schaal van 1 tot 7, waarbij geldt 1 = volledig mee oneens en 7 = volledig mee eens. Omcirkel het antwoord van je keuze.

2. Deze tekst bevat veel moeilijke zinnen.

ONEENS 1 2 3 4 5 6 7 EENS

3. Deze tekst is poëtisch.

ONEENS 1 2 3 4 5 6 7 EENS

4. Deze tekst bevat veel dubbelzinnigheden.

ONEENS 1 2 3 4 5 6 7 EENS

5. Deze tekst is prettig om te lezen.

ONEENS 1 2 3 4 5 6 7 EENS

6. Ik vond deze tekst moeilijk om doorheen te komen.

ONEENS 1 2 3 4 5 6 7 EENS

7. De schrijfstijl van deze tekst is onprettig.

ONEENS 1 2 3 4 5 6 7 EENS

8. Deze tekst komt onnatuurlijk en gekunsteld over.

ONEENS 1 2 3 4 5 6 7 EENS

9. De zinnen in deze tekst zijn te lang.

ONEENS 1 2 3 4 5 6 7 EENS

10. Deze tekst kan gekenmerkt worden als literatuur.

ONEENS 1 2 3 4 5 6 7 EENS

11. Er worden in deze tekst onlogische gedachtesprongen gemaakt.

ONEENS 1 2 3 4 5 6 7 EENS

12. Deze tekst bevat vreemde woordkeuzes.

ONEENS 1 2 3 4 5 6 7 EENS

13. Deze tekst vormt een eenheid.

ONEENS 1 2 3 4 5 6 7 EENS

14. In deze tekst lijken de woordkeuze en de formulering belangrijker dan de inhoud.

ONEENS 1 2 3 4 5 6 7 EENS.

15. Deze tekst is boeiend.

ONEENS 1 2 3 4 5 6 7 EENS

16. Deze tekst is beeldend geschreven.

ONEENS 1 2 3 4 5 6 7 EENS

17. Deze tekst leest vlot.

ONEENS 1 2 3 4 5 6 7 EENS

18. Deze tekst is eentonig.

ONEENS 1 2 3 4 5 6 7 EENS

19. De auteur had hetzelfde kunnen zeggen in veel minder woorden.

ONEENS 1 2 3 4 5 6 7 EENS

20. De zinnen in deze tekst zijn te kort.

ONEENS 1 2 3 4 5 6 7 EENS

21. Deze tekst is ondubbelzinnig.

ONEENS 1 2 3 4 5 6 7 EENS

22. De verschillende paragrafen in deze tekst staan in een logische volgorde.

ONEENS 1 2 3 4 5 6 7 EENS

23. Deze tekst heeft een duidelijk verhaal.

ONEENS 1 2 3 4 5 6 7 EENS

24. Deze tekst leest stroef.

ONEENS 1 2 3 4 5 6 7 EENS

25. Deze tekst heeft een prettig ritme.

ONEENS 1 2 3 4 5 6 7 EENS

26. Deze tekst is warrig.

ONEENS 1 2 3 4 5 6 7 EENS

27. Deze tekst bevat vreemde zinsconstructies.

ONEENS 1 2 3 4 5 6 7 EENS

28. Deze tekst heeft een rommelige structuur.

ONEENS 1 2 3 4 5 6 7 EENS

29. De auteur heeft woorden gebruikt die ik normaal gesproken niet op die manier zou gebruiken.

ONEENS 1 2 3 4 5 6 7 EENS

30. Deze tekst is saai.

ONEENS 1 2 3 4 5 6 7 EENS

31. Deze tekst wordt gekenmerkt door ambtelijk taalgebruik.

ONEENS 1 2 3 4 5 6 7 EENS

32. De verschillende paragrafen in deze tekst hebben weinig met elkaar te maken.

ONEENS 1 2 3 4 5 6 7 EENS

33. Deze tekst is suf.

ONEENS 1 2 3 4 5 6 7 EENS

34. Deze tekst bevat ouderwets taalgebruik.

ONEENS 1 2 3 4 5 6 7 EENS

35. Deze tekst is geschreven door een amateur.

ONEENS 1 2 3 4 5 6 7 EENS

36. Deze tekst leest moeilijk.

ONEENS 1 2 3 4 5 6 7 EENS

37. Deze tekst is afwisselend.

ONEENS 1 2 3 4 5 6 7 EENS

38. Deze tekst bevat geen duidelijk verhaal.

ONEENS 1 2 3 4 5 6 7 EENS

39. Het ritme in deze tekst is niet natuurlijk.

ONEENS 1 2 3 4 5 6 7 EENS

40. De tekst heeft een duidelijke opbouw.

ONEENS 1 2 3 4 5 6 7 EENS

41. Deze tekst bevat geen zin teveel.

ONEENS 1 2 3 4 5 6 7 EENS

42. Deze tekst is grappig.

ONEENS 1 2 3 4 5 6 7 EENS

C.

Geef aan hoe gevarieerd je het gebruik van de verschillende woordsoorten vindt, waarbij 1 = heel eentonig, de auteur gebruikt steeds dezelfde woorden, en 7 = heel gevarieerd, de auteur gebruikt een heleboel verschillende woorden. Het gaat hierbij om je gevoel over deze tekst, je hoeft de woorden niet daadwerkelijk te gaan tellen.

Om je te helpen is hier een voorbeeld van de woordsoorten die bedoeld worden:

"*Mijn*	*nieuwe*	*huisgenoot*	*heeft*	*gisteren*	*rotte*	*appels*	*gekocht*."
	BIJV.NW.	ZELFST.NW.	WW.	BIJW.	BIJV.NW.	ZELFST.NW.	WW.

43. Hoe beoordeel je de variatie in werkwoorden?

(Bijv. "heeft", "gekocht")

HEEL EENTONIG 1 2 3 4 5 6 7 HEEL GEVARIEERD

44. Hoe beoordeel je de variatie in zelfstandig naamwoorden?

(Bijv. "huisgenoot", "appels")

HEEL EENTONIG 1 2 3 4 5 6 7 HEEL GEVARIEERD

45. Hoe beoordeel je de variatie in bijvoeglijk naamwoorden en bijwoorden?

(Bijv. "nieuwe", "gisteren", "rotte")

HEEL EENTONIG 1 2 3 4 5 6 7 HEEL GEVARIEERD

D.

Geef aan of, en zo ja in welke mate, je het eens bent met de volgende stellingen. Omcirkel het antwoord van je keuze en geef een voorbeeld als je 'ja, enkele' of 'ja, veel' geantwoord hebt.

46. Deze tekst bevat fouten in zinsbouw, woordvolgorde en grammatica.

Bijvoorbeeld: "*<u>Hun</u> komen morgen op de koffie*."

NEE JA, ENKELE JA, VEEL

GEEF EEN VOORBEELD: ....

....

47. Deze tekst bevat fouten in woordkeuze en woordgebruik.

Bijvoorbeeld: "*Ik ga mee naar het feest, <u>mits</u> ik niet mag van mijn moeder. Ik zal haar meteen <u>optelefoneren</u>.*"

NEE JA, ENKELE JA, VEEL

GEEF EEN VOORBEELD: ....

....

48. Deze tekst bevat spellingsfouten.

Bijvoorbeeld: "*Dat <u>vindt</u> ik ook*."

NEE JA, ENKELE JA, VEEL

GEEF EEN VOORBEELD: ....

....

49. In deze tekst komt foutief gebruik van leestekens voor.

Bijvoorbeeld: "*Is het waar dat er in Nederlandse bodem bewegende aardschollen zijn waargenomen*."

NEE JA, ENKELE JA, VEEL

GEEF EEN VOORBEELD: ....

....

E.

50. Geef deze tekst een cijfer op een schaal van 1 tot 10: _______________

Kun je zeggen waar dit cijfer vooral op gebaseerd is?

F.

51. Zijn er verder nog dingen aan deze tekst die je zijn opgevallen maar die niet in deze vragenlijst aan bod gekomen zijn?

52. Heb je verder nog opmerkingen over deze tekst of over deze vragenlijst?

53. Had je deze tekst al eens eerder gelezen?

Ja   Nee, maar wel iets wat er sterk op lijkt   Misschien   Nee

Hartelijk dank voor je medewerking!

C Answers to Question 1 of the questionnaire

This question asked participants for their first opinion of the text.

Table C1

No.	Condition	Q1
1	No ZL	Korte zinnen. Loopt meestal vloeiend, on-nederlands. Erg geforceerd lijkt het.
2	No ZL	Eenvoudige, directe stijl. Aparte woordkeuze, zoals ‘tetteren’, ‘tieten’, ‘ik draai me naar’. Overgangen tussen zinnen zijn niet altijd even vloeiend. Wél duidelijk en goed/snel leesbaar.
3	No ZL	Een rommelige, enigszins onsamenhangende tekst. Het verhaal wordt vrij simpel verteld, met korte zinnen. Het springt een beetje van de hak op de tak, het lijken losse gedachten. De ene keer wordt er gesproken over ‘mevrouw’, de andere keer over ‘Tineke’. Het zijn een beetje kromme zinnen. De woorden die gebruikt zijn, zijn duidelijk en absoluut niet moeilijk.
4	No ZL	Over het algemeen zou ik niet meteen zeggen dat het door een beginnende vertaler geschreven is; het lijkt meer dan beginnend. De stijl is makkelijk. De schrijver vertaalt duidelijk het gevoel van de ik-persoon: in de eerste zin al. Ik verbaasde me over de woordkeus van ‘overhevelen’, iets boven het midden op pag. 2. Ik hoor mensen niet vaak dat woord gebruiken. Verder is het taalgebruik overmatig makkelijk, namelijk.
5	No ZL	Hele korte zinnen, soms iets te kort naar mijn mening. Ook al is het wel een goede weergave van hoe de taal er in gedachtes uitziet, vind ik dat lange, verhalende zinnen meer bij dit soort teksten passen. Dialogen zijn soms onduidelijk in de zin dat het niet duidelijk is wie wat zegt. Zoals op pag. 3: ‘weet je wat... aan je voorwendselen’. De zin daarna (Ze kijkt me...) kan beter op de volgende regel, want er wordt geschakeld tussen de personen betrokken bij de dialoog.
6	No ZL	De zinnen zijn afwisselend kort en lang. De zinnen komen vaak gekunsteld over, zoals in "Alleen al … een Aziatische." De stijl en het woordgebruik is afwisselend informeel en formeel.
7	No ZL	Sommige passages hebben een sarcastische en/of spottende ondertoon. Ter illustratie: "Zul je altijd zien: zit zo’n vrouw eens in een boek te lezen en hup, ze heeft ineens wat in haar knieën." De verteller is of met het verkeerde been uit bed gestapt, óf hij is enigszins verbitterd. Daarnaast worden zinnen niet gewoonlijk geconstrueerd. Er valt niet één duidelijke lijn in te ontdekken.
8	No ZL	Een paar keer wordt ‘welk’ gezegd, terwijl voor mijn gevoel ‘wat voor’ beter past. Woordvolgorde gek: "dat ik er pijn van in mijn zij had" beter is -> "dat ik er pijn in mijn zij van had". "Ik wil zien" -> "ik zou wel eens willen zien". "Dat lijkt het niet" -> "daar lijkt het niet op". Alles is in dezelfde stijl geschreven, gedacht en van 1 persoon. "Kom hier" (blz. 3) -> "geef eens hier" (bijv.) Sommige constructies zouden wij niet gebruiken. In plaats daarvan begruiken we zinnen die een beetje gek zijn, soort uitdrukkingen. "Wat heb je met je knieën" -> "aan je knieën".
9	No ZL	Af en toe redelijk ingewikkelde woorden in een anders simpele tekst. Het ‘fresh-prince grapje’ zal voor veel Nederlanders onbekend zijn. Sommige uitdrukkingen worden nauwelijks in het Nederlands gebruikt ("Veer ik kort op achter mijn bureau", p.3) ("In mijn werk... kort te houden", p.2). Sommige woorden worden nauwelijks in het NL gebruikt ("voorwendselen", p. 3)
10	No ZL	De tekst is geschreven in de ik-vorm en op sarcastische wijze. Het zijn de gedachten van iemand want soms is het taalgebruik ordinair en plat. Uit de paar zinnen die hij zegt tegen mensen is hij echter ook vervelend. Het lijkt een kort verhaal die matschappij-kritisch is tegenover arbeidsongeschikten.
11	No ZL	Het is vanuit een ik-perspectief geschreven. Het is goed te begrijpen al is de woordkeuze soms wat moeilijker/deftiger. Ook viel de referentie aan fresh prince me op. Misschien weet niet iedereen meer wie dat is of waar dat over gaat. De paragrafen lopen niet altijd in elkaar over, het vormt niet een passend geheel.
12	No ZL	Regelmatig gebruik van woordherhalingen zoals getetter en rapport op de laatste bladzijde. Soms zijn de zinnen onoverzichtelijk. Bijv. "…zou ik Tineke nu liever snel weer hier willen terugzien"
13	No ZL	Het is goed Nederlands. Niet erg zakelijk taalgebruik, maar wel goed te begrijpen. Het lijkt heel erg op spreektaal, soms zit er geen werkwoord in de zin. De zinnen zijn vrij kort en er is veel dialoog. Toch is het lastig uit de tekst op te maken wat voor situatie de hoofdpersoon zich nu in bevindt. Er woorden vooral veel gedachtes verwoord.
14	No ZL	Veel gebruik van ‘zou’ en ‘zal’, zoals in het Engels ook veel gedaan wordt. De zinnen komen vaak wat vreemd over, en sommige zoals "Dat lijkt het niet" kloppen niet. De stijl is verwarrend. Soms lijkt het heel serieus en zakelijk te zijn en soms juist precies het tegenovergestelde. Het woord

		'bewuste' komt een aantal keer terug. Het is wel Nederlands maar zo vaak wordt dat niet gebruikt dus dat komt wat vreemd over.
15	No ZL	Leest gemakkelijk, maar het klinkt een beetje kunstmatig, over-correct.
16	No ZL	Het leest makkelijk door. Niet te veel moeilijke woorden. Dagelijks taalgebruik. Beetje korte zinnen af en toe -> iedere alinea begint met een korte zin.
17	No ZL	Zinnen sluiten niet mooi op elkaar aan met bijvoorbeeld een signaalwoord. Het zijn veel korte zinnen die ook met een bepaald signaalwoord aan elkaar verbonden hadden kunnen worden. Soms woorden die we in het Nederlands amper gebruiken, zoals 'voorwendselen' en 'getetter'. Er is niet veel afwisseling in woordkeuze, 'meneer', 'mevrouw' en 'getetter' komen vaker voor en hadden ook met een ander woord omschreven kunnen worden, bijv. 'geklets'.
18	No ZL	In principe is het een leesbare tekst, met grammaticale zinnen. Door de woordkeuze en de volgorde van bepaalde zinsdelen doet het soms wat kinderlijk of vreemd aan: het is grammaticaal juist, maar het zijn woordvolgordes die wat minder vaak voorkomen en het is een redelijk simpel vocabulaire, met minder afwisseling van bepaalde woorden en zinsconstructies dan ik gewend ben van een oorspronkelijk Nederlandse tekst.
19	No ZL	Veel inner speech, dus niet uitgesproken gedachten. Lange zinnen. Duidelijke alineaverdeling.
20	No ZL	De stijl komt in sommige stukken nogal stijf over, hoewel sommige stukken juist een redelijk informeel onderwerp hebben. Er worden vaak woorden herhaald, zoals 'Tineke' in het midden van bladzijde 2, of 'rapport', op bladzijde 3. De zinnen lijken op de een of andere manier niet goed in elkaar over te lopen.
21	No ZL	Er wordt alleen gepraat vanuit de ik-persoon, die redelijk negatief en zakelijk in het leven staat. Hij/zij lijkt een behoorlijke pessimist, met een goede achtergrond gezien de grote woordenschat en het taalgebruik.
22	No ZL	Korte zinnen; makkelijke woorden; verkeerde alinea's
23	No ZL	De tekst is moeilijk te volgen aangezien sommige zinnen grammaticaal niet helemaal kloppen, waardoor je een bepaalde zin twee of meer keer moet lezen voordat je de zin begrijpt. Dit heeft als gevolg dat je moeite krijgt met het begrijpen van het verhaal als één geheel.
24	No ZL	Ik vond het heel vervelend lezen. Weinig samenhang door verbindingswoorden tussen de zinnen. Het is heel letterlijk vertaald denk ik.
25	No ZL	Gebruik van de werkwoorden is redelijk onnatuurlijk, de zinnen zijn soms vreemd opgebouwd en net iets te kort of iets te lang, het is stroef om te lezen. De woordkeuze is wel aardig gevarieerd.
26	No ZL	Het leest niet lekker door. Heel kort taalgebruik. Soms klopt de zinsopbouw niet helemaal. Ook moet je soms een paar keer lezen voordat je het goed begrepen hebt.
27	No ZL	Het leest erg onrustig, het is een informele tekst en woordkeuze. Er wordt veel geciteerd.
28	No ZL	Naar mijn idee is deze tekst letterlijk vertaald. Het leest niet echt fijn en sommige zinnen lopen niet goed of sluiten niet goed op elkaar aan. Spreektaal en zakelijke taal worden door elkaar heen gebruikt.
29	No ZL	Ik had het idee dat het nogal letterlijk uit het Engels is vertaald, waardoor het moeilijk was om de tekst te begrijpen. Dat het letterlijk vertaad is zie je soms ook aan de woordkeuze -> "zat" van dat getetter.
30	No ZL	Redelijk goed Nederlands met af en toe niet helemaal kloppende woorden. Soms iets te lange zinnen of moet er ergens een komma tussen. Vreemde overgangen naar ineens een ander onderwerp.
31	No ZL	Het was in hele makkelijke taal geschreven, geen lastige woorden.
32	No ZL	Het loopt niet echt, allemaal losse zinnen di geen geheel vormen. "Precies zes hele" vind ik niet mooi (midden blz. 1). Woordkeuze loopt door elkaar: beschaafd en netjes, dna ineens niet meer, taalgebruik vind ik niet mooi.
33	No ZL	De schrijfstijl lijkt wat losstaand van gevoel, wat feitelijker, iets minder natuurlijk. Ook de woordkeuze lijkt wat minder natuurlijk. Dit gevoel had ik vooral op de eerste bladzijde.
34	No ZL	Ik kom niet zozeer vreemde woorden of vormen tegen, maar de tekst leest niet echt lekker; het verhaal loopt niet echt goed door. Woordkeuze zie ik niet echt gekke dingen, maar toch is het taalgebruik vreemd.
35	No ZL	Het verhaal loopt niet echt lekker. Er zijn woorden gebruikt die ik zelf niet zo zou gebruiken, omdat ze tamelijk formeel overkomen. Ook is er een beetje een krampachtige overgang van korte en lange zinnen. Ook de woordvolgorde/zinsdeelvolgorde is hier en daar lastig om vlot te lezen.
36	No ZL	Ik heb moeite met het begrijpen van enige alinea-overgangen, vooral wanneer er een nieuw onderwerp aangehaald wordt. De overgangen zijn plotseling en niet altijd even duidelijk, waardoor ik de zin een tweede keer moet lezen. Ze in enkelvoud, de volgende regel in meervoud. Sommige zinnen missen werkwoorden, andere bevatten verwijswoorden die nergens naar schijnen te verwijzen.
37	No ZL	Droge omschrijving van omgeving; woordkeuze niet altijd logisch, soms passen de woorden niet bij de zin.
38	No ZL	Het is nogal een opdreuning van zinnen achter elkaar, het leest niet lekker. De woordkeuze is soms best grof.
39	No ZL	Zinnen worden onregelmatig afgebroken: soms zijn zinnen te kort, soms zijn zinnen erg lang. Veel woordherhalingen binnen zinnen. Veel omschrijvende zinnen ipv concrete zinnen.

40	No ZL	Ik vind het niet heel soepel lezen. Het komt veel voor dat de vertaler behoorlijk letterlijk heeft vertaald, en sommige zinstructuren komen wat stijfjes over in het Nederlands. Er zijn weinig verbindingswoorden om het verhaal vloeiend te laten lopen en soms is het gebruik van tense een beetje verwarrend. Soms 'had' waar een andere verledentijdsvorm beter had gepast etc.
41	No ZL	Vrij groffe tekst, beeld vanuit de auteur, korte zinnen.
42	No ZL	Het begin kon ik niet goed volgen, het leek alsof de vertaler eerst nog in het verhaal moest komen want naarmate de tekst vordert wordt het beter leesbaar. Er zijn woorden die herhaald worden, zoals 'getetter', die soms storend kunnen zijn. De stijl vind ik nogal kortaf, zakelijk, to the point, maar dat kan de stijl zijn van de originele tekst.
43	No ZL	Makkelijk taalgebruik, tieten in plaats van borsten bijvoorbeeld. Korte zinnen en makkelijk te lezen.
44	No ZL	De stijl is makkelijk, het klopt grammaticaal wel altijd maar er waren veel zinnen die ik net anders zou doen. Ook viel het me op dat ze meneer en mevrouw gebruikte ipv man en vrouw.
45	No ZL	Korte zinnen, grappig.
46	No ZL	Woordkeuze is simpel, stijlvorm richt zich op korte zinnen, om de toon van het verhaal beter weer te geven. Taalgebruik is vrij elementair zonder opsmuk of verbloeming, "in your face".
47	No ZL	Formeel taalgebruik maar soms ook informeel. De schrijfstijl is alsof iemand een verhaal verteld, dus alsof iemand zelf spreekt.
48	No ZL	Het gebruik van tijd vind ik niet natuurlijk overkomen. Er wordt veel in de voltooid verleden tijd geschreven terwijl dat niet nodig is (OVT zou ook goed zijn). En soms wordt er gewisseld van verleden naar tegenwoordige tijd. Soms worden er ook opeens dure woorden gebruikt, "welgevallen" blz. 1 is een voorbeeld daarvan. Die komen in mijn ogen een beetje uit het niets.
49	No ZL	Erg korte zinnen, die onprettig in elkaar overlopen. Vaak is het lastig te volgen waar een nieuwe alinea over gaat (gaat ineens over iets anders). De tekst heeft een informele stijl, maar is niet in spreektaal geschreven: vaak worden er constateringen gedaan, die in spreektaal niet in die vorm terug te vinden zijn.
50	No ZL	Wat mij opvalt, is dat ik het lastig lezen vind. Qua taal is het goed te begrijpen, ik kan me ook inleven in de ik-persoon, maar soms m.n. aan het begin, gaat het snel van het één over in het ander, waardoor ik terug moet gaan in het verhaal (wat mij overigens niet heel veel meer duidelijkheid oplevert). Vooral dat het echt van het één in het ander overgaat vind ik lastig. Qua woordkeuze, vorm en taalgebruik vind ik het prima; echter zit daar wellicht wel een te groot verschil tussen stijl enerzijds en de andere aspecten anderzijds.
51	No ZL	Vreemd taalgebruik zo nu en dan. Ik kan nauwelijks merken dat de tekst uit een andere taal vertaald is. Ook lijken sommige woorden aan het Nederlands aangepast te zijn. De namen 'Tineke' en 'els' ogen bijvoorbeeld wel erg Nederlands.
52	No ZL	Wat mij direct opvalt is het gebruik van korte zinnen, in de meeste gevallen erg gemakkelijk. Er wordt gekozen voor een stijl waarin erg korte alinea's gebruikt worden. Er worden veel woorden gebruikt als zal en zou. Vrij weinig is met zekerheid geschreven.
53	No ZL	De zinnen lopen niet lekker. Het is vrij letterlijk vertaald. Woordkeuze had bij veel woorden net iets anders moeten zijn om het lopend te maken. De stijl is niet erg coherent. De ene keer is de tekst heel formeel, dan weer informeel.
54	No ZL	Ik vond de tekst niet zo lekker lopen en zeker in met woordvolgorde in bepaalde zinnen had ik een beetje moeite. Dit maakte dat ik er niet echt soepel doorheen kon lezen.
55	No ZL	De woordkeuze lijkt hier en daar niet al te logisch. Het lijkt alsof de schrijver wel kennis heeft van bepaalde Nederlandse zegswijzen maar hier ook fouten in maakt. - De stijl is buitengewoon informeel met hier en daar een moeilijk woord als "welgevallen". Verder mist er terminologie die je wel zou verwachten. 'Haar soort mensen' had wellicht een term als hypochonders oid moeten zijn. - Het taalgebruik is hier en daar vrij plat en grof.
56	No ZL	Vreemd opgebouwde zinnen; vreemde uitspraken; veel "omgekeerde" zinnen; rare vraagstellingen; toekomstige/verleden tijd raar afgewisseld
57	No ZL	Slechte verwijzingen, wisseling van 'tijden', weinig structuur dmv verbindingswoorden
58	Original	Vanuit de gedachten van een man, kritisch, makkelijk leesbaar, spreektaal, geen moeilijke woorden, korte spanningsboog, verhalend, soms metaforisch taalgebruik.
59	Original	Normale Nederlandse tekst met een paar opvallendheden: "zitten gaan" bovenaan bladzijde 3 bijvoorbeeld en een "tijdelijk openstaande deur" klinkt ook raar.
60	Original	Veel subjectief/informeel taalgebruik (m.n. wat betreft woordkeuze), eenvoudige, korte zinnen. Weinig bijzinnen en (vrijwel) geen lange/moeilijke woorden. Stijl is direct.
61	Original	Er zitten een aantal lange zinnen in, die je 2x moet lezen om het goed te begrijpen. Ook zijn er woorden uit het Engels vertaald die Nederlanders nauwelijks meer gebruiken. Soms staat een werkwoord niet op de plek waar je deze verwacht.
62	Original	Goed geschreven, maar er werden opvallende idiomen gebruikt af en toe, zeker voor interne monologen. Er worden af en toe constructies en idiomen gebruikt die niet vaak voorkomen in spreektaal, het is best wel geschreven in schrijfstijl, bijna romanstijl. Eén keer ben ik een fout tegengekomen, verder is het prima geschreven (de fout is de verkeerde tijd in 'gebeurt' op de eerste pagina), het leest vlot en makkelijk.

63	Original	De tekst is makkelijk te lezen. Makkelijke woorden, niet al te ingewikkelde zinnen. Maar sommige alinea-stukjes zijn een beetje van de hak op de tak.
64	Original	Het lijkt me een tekst uit een roman, geschreven met het doel te amuseren; sterk verhalend, informeel (behalve de letterlijk geciteerde uitspraken), taalgebruik. Af en toe een wat formele uitdrukking en soms erg informeel (forse tieten, me eigen)
65	Original	Veel ‘vlot’ taalgebruik. Niet al te lange zinnen. Woordkeuze lijkt gericht te zijn op een doelgroep zonder jargon.
66	Original	Vooral in het begin staan er ingewikkelde zinnen in. Het loopt niet heel vloeiend.
67	Original	Bij de meeste teksten die uit het Engels naar het Nederlands vertaald zijn kun je de Engelse zin nog wel makkelijk herleiden. Bij deze niet. Bij deze tekst is er volgens mij erg gelet op de woordkeuze, er zijn bijna geen woorden gebruikt die in het Engels hetzelfde zijn als in het Nederlands. De zinnen zijn ook leesbaar, korte en lange zinnen wisselen elkaar goed af.
68	Original	Mijn eerste indruk was dat het er goed uitziet, ik vind de opbouw van de zinnen goed en ook de woorden die gebruikt zijn, zijn goed gekozen. Ik denk ook wel dat de stijl hetzelfde is als de oorspronkelijke tekst, of de vertaler heeft de stijl heel erg veranderd maar dat geloof ik niet.
69	Original	De zinnen zijn niet heel lang. In sommige gevallen had ik bijvoorbeeld een komma gezet, waar deze in de tekst de zin wordt beëindigd met een punt.
70	Original	Woordkeuze is wat grof. Voor meer structuur in de tekst en om het wat leesbaarder te maken had de vertaler meer gebruik van leestekens kunnen maken. Het taalgebruik is cynisch.
71	Original	Ik vind de tekst goed te begrijpen, aardig goed vertaald. Het loopt nog niet helemaal vloeiend als je het hardop zou voorlezen. Af en toe is een woord minder goed vertaald, klopt het niet in de context.
72	Original	Het verhaal is een weergave van gedachtes. Soms worden lange, komische zinnen gebruikt en daarnaast korte zinnen die niet compleet zijn, maar een gedachte weergeven. Bijvoorbeeld in: ‘Precies wat ik dacht. Een vrouw.’ Dit zou ik meer spreektaal noemen dan schrijftaal. Daarentegen worden soms ook weer lange, complexere zinnen gebruikt met veel bijwoorden en bijvoeglijk naamwoorden die een humoristische sfeer creeëren.
73	Original	Humoristisch geschreven. Bevat veel sarcasme. Soms is het wat lastig te volgen doordat het verhaal een beetje "van de hak op de tak" springt. Daarnaast zijn sommige zinnen niet heel prettig om te lezen door een vreemde woordvolgorde of woordkeuze.
74	Original	De vertaling is soms best wel plat opeens, dat over tamelijk forse tieten zou ik nooit gebruiken. Ook wordt de mevrouw die last heeft telkens anders genoemd, eerst gewoon mevrouw, dan Tineke en dan mevrouw Burgers. Er zijn geen andere vrouwen in het verhaal, maar anders zou dat verwarrend kunnen zijn. Ook zijn de zinnen heel kort en best veel zinnen met alleen een lidwoord en een zelfstandig nw, dat leest niet echt lekker. Ook vind ik de ik-persoon veel te sarcastisch en gemeen, maar dat is meer de inhoud van de tekst. En het zijn ook vaak gedachtes van de ik-persoon, maar ze worden wel indirect opgeschreven dus het komt heel aserieus en slordig over.
75	Original	Woordkeus is niet altijd juist gekozen, soms nogal ongepast taalgebruik (of echt een verkeerd woord) dat door andere woordkeus beter had kunnen zijn. Stijl is niet zakelijk, soms ook ongepast. Verkeerde grammatica.
76	Original	Geschreven vanuit 1e persoon, verwoordt letterlijk de gedachten van de hoofdpersoon wat blijkt uit het platte taalgebruik. Makkelijk te lezen / volgen.
77	Original	De tekst bevat veel sarcasme. Veel dingen worden ook overdreven of met beeldspraak uitgedrukt. De woordkeuze is overwegend simpel. Korte zinnen wortden afgewisseld met hele lange zinnen.
78	Original	Het komt op mij vrij cynisch over; Het lijkt op een fragment uit een roman; de woordkeuze vind ik soms een beetje te zakelijk; de manier waarop iemand iets zegt wordt expliciet beschreven.
79	Original	De woordkeuze oogt wat formeel.
80	Original	De tekst leest niet zo vlot als zou kunnen. Dat komt doordat naar mijn idee sommige zinnen een beetje omslachtig zijn verwoord: niet helemaal lekker lopen. De betekenis is duidelijk, maar toch moest ik sommige zinnen of woorden (materies) 2 keer lezen. Vooral de zinsopbouw is verrassend af en toe.
81	Original	Gek geformuleerde zinnen. Tevens zijn alle zinnen op dezelfde manier geformuleerd. Amper bijzinnen. Korte zinnen.
82	Original	Een aantal vaste uitdrukkingen zoals "dat interesseert me geen reet" kloppen niet. Daarnaast zijn bepaalde zinsconstructies rechtstreeks overgenomen uit het Engels: "communicatief niet vaardig houden zij…" Verder ook juist goed gebruik van uitdrukkingen op andere plaatsen. Sommige zinnen moest ik voor mijn begrip ervan herlezen.
83	Original	Concrete tekst; opmerkelijk veel woorden die niet dagelijks worden gebruikt; spelen met de taal; informeel taalgebruik
84	Original	Het taalgebruik is voornamelijk informeel, waardoor het een stuk gemakkelijker wegleest.
85	Original	Ik vind de woordkeuze soms wel wat aan de moeilijke kant. De tekst is wel echt bedoeld voor mensen met ervaring met literatuur en die een grote woordenschat hebben. - Taalgebruik is net als woordkeuze gevorderd. - Verder wisselt het heel erg tussen lange en korte zinnen.
86	Original	Gedetailleerde stijl. Veel verschillende tijdsvormen door elkaar; verleden tijd, voltooide tijd, tegenwoordige tijd en toekomende tijd. Veel bijvoegelijke naamwoorden.

87	Original	De woordkeuze is soms een beetje eigenaardig, spreekwoorden die deze man schrijft bestaan ook niet allemaal. De vorm is anders, soms verkeerde woordvolgorde.
88	Original	Woordvolgorde soms niet logisch. Woordkeuze soms niet passend in de context en te letterlijk vertaald.
89	Control	Ik vind de woordvolgorde in de tekst soms redelijk geforceerd en ik kan er ook duidelijk aan zien dat de tekst oorspronkelijk Engels was. Wat mij verder opvalt is dat er altijd 'zij' en 'mij' wordt gebruikt, maar eigenlijk nooit 'ze' en 'me', wat in het Nederlands redelijk ongebruikelijk is volgens mij. Daardoor komt de tekst een beetje afstandelijk op mij over.
90	Control	Onsamenhangende tekst; gedachten/gebeurtenissen volgen elkaar snel op, er wordt ook spreektaal opgeschreven, sowieso hele informele stijl. Soms plotseling formele woorden tussendoor, die niet zo passen.
91	Control	Er wordt een aparte stijl gebruikt met soms vreemd gekozen woorden, een beetje ouderwets of complex soms. Ook klopt de vorm niet altijd zoals combinaties bij een bepaald werkwoord. Het komt over het algemeen op me over als taalgebruik van een niet-moedertaalspreker van het Nederlands.
92	Control	Er worden veel woorden gebruikt die een moedertaalspreker van het Nederlands niet zo snel zou gebruiken. Af en toe zijn er ook echt verkeerde vertalingen. Het valt verder op dat er veel letterlijke vertalingen zijn die je in het Nederlands nooit op die manier zou horen.
93	Control	(1) Wat me opviel is dat persoonlijk voornaamwoorden vaak onnatuurlijk overkomen. Er wordt alleen maar gebruik gemaakt van 'jij', 'jou', 'mij' en nooit 'je' of 'me' wat mij in sommige gevallen natuurlijker lijkt. (2) Zo nu en dan worden er ver-Engelste woorden gebruikt, bijv. corporaties.
94	Control	- Soms te weinig overgang van de ene naar de andere zin, waardoor het twee compleet losse nieuwe zinnen lijken. - Zinnen beginnen met 'alsof' of 'maar' of 'al zou ik', wat je meestal in een bijzin zou verwachten. - Rare woordkeuzes: "het zou me niks verbazen als dat plaatsvindt" ipv "gebeurt". Dat je gewoon bent aan jouw voorwendselen (ipv gewend geraakt aan). Neem een stoel (te letterlijk take a seat, in NL: ga zitten). - Persoonlijk voornaamwoorden soms onduidelijk. bijv: "Communicatief houden zij zich niet staande" etc. - Woordvolgorde: ik zou zeggen: "Dat ik er pijn in mijn zij van had" ipv "mijn van in mijn zij had".
95	Control	Het lijkt alsof woorden en structuren waar mogelijk letterlijk zijn vertaald. De Engelse stijl van spreken lijkt ook letterlijk overgenomen.
96	Control	Beetje gekunsteld, stijl is wat stroef, veel gebruik van gemarkeerd "zij" i.p.v. "ze". Taalgebruik is ietwat generiek tot soms weinig specifiek, wat een nonchalante, onverschillige toon veroorzaakt.
97	Control	Korte zinnen, soms zinnen uit één enkel woord, tegenwoordige tijd, een erg letterlijke en subjectieve beschrijving wat er in het hoofd van de hoofdpersoon gebeurt. Een beetje stream-of-conciousness-acthig? Geen openende apostrofs bij spraak.
98	Control	De tekst is heel lastig te volgen door het taalgebruik. Ik kan me voorstellen dat de originele Engelse tekst een stuk leesbaarder is. Er wordt naar mijn idee vaak heel archaïsch vertaald, zowel op woord- als op zinsniveau, terwijl dit niet altijd nodig lijkt te zijn. Op andere plaatsen is de vertaling dan weer te letterlijk ("kenmerkend voor haar soort", "opwekkend", "imiteren dat ik aan het lezen ben"). Verder gaat het bij de dialoog fout met aanhalingstekens. Is "me eigen" een fout van de vertaler, of probeert hij een soort stijlelement toe te voegen?
99	Control	De tekst leest niet echt makkelijk. Beetje ouderwets en ongewoon woordgebruik en ook de stijl is apart, wel correct maar niet gebruikelijk. Vaak komen de Engelse constructies naar voren, alsof ze iets te letterlijk zijn vertaald of om dicht bij de oorspronkelijke tekst te blijven. Verder komt de tekst wat oubollig over doordat veel "zij" wordt gebruikt in plaats van "ze", wat de tekst ook een formeel karakter geeft. Soms is de tekst wat simpel ("hij vraagt of zij snel weer terugkeert"), wat in het Engels een geaccepteerde zin is, maar in het Nederlands dus wat simpel overkomt. Ook viel de zin "ik wend me eigen niets voor" op, redelijk spreektaal, wat misstaat in zo'n degelijke tekst.
100	Control	Door de korte alinea's is de tekst goed te volgen. Het taalgebruik van de ik-persoon is 'lastiger' dan je normaal gesproken zou verwachten. Vanuit de ik-persoon krijg je in dit verhaal de situatie waarin hij/zij zich bevindt te horen en de gedachten die deze ik-persoon heeft. Bij de gedachten van een persoon zou ik niet verwachten dat er een woord als 'ventileren' gekozen wordt in plaats van het veel gebruikelijkere 'luchten' ("om deze werkdag te overleven moest ik het zooitje flink ventileren / luchten")
101	Control	Het verhaal is in de ik-vorm geschreven. Er wordt veel gebruikgemaakt van bijvoegelijk naamwoorden en bijzinnen. Het zal vast aan de vertaler liggen, maar in de tekst staan woorden of tekstelementen die een persoon die Nederlands spreekt doorgaans niet zo snel zhou gebruiken. Die woorden komen nogal bekakt of gewoon raar; het voelt alsof het er anders had moeten staan om natuurlijker te kunnen komen.
102	Control	Doordachte tekst. Er zit meer achter dan je in eerste instantie zou verwachten. Veel gebruik van 'zij' waar ik zelf eerder 'ze' zou gebruiken.

103	Control	Aparte woordkeuze, woorden die ik zelf niet als vertaling zou gebruiken. De tekst voelt onnatuurlijk aan. Heel veel korte zinnen. Ik kan me niet inleven in de hoofdpersoon door de manier van schrijven.
104	Control	De teks is geschreven in 'denktaal', op een manier die je niet snel zult lezen in een boek of krant. Dat over stijl. De woordkeuze gaat hier in mee, het is losjes geschreven en soms moeilijk te volgen.
105	Control	Heel veel woorden zijn vertaald met woorden die in het Nederlands weinig worden gebruikt. Erg letterlijke vertaling van woorden. De grammatica was daarentegen wel vrij goed. Ik vond de uiting 'me eigen' interessant. Komt dit door invloed van een dialect waar ze Nederlands leert (ik hoor in Brabant deze uiting wel vaker) of is het gewoon slecht vertaald.
106	Control	Er wordt redelijk wat beeldspraak gebruikt, en ook sarcasme waarbij iets anders gezegd wordt dan echt bedoeld wordt. Ook staan er vaak woorden in die ikzelf in ieder geval niet snel zou gebruiken, ookal zijn het geen moeilijke woorden ("billijkheid", "zielenrust", "voorwendselen"). Verder is het op een spreektaal-achtige manier geschreven.
107	Control	aparte/complexe woordkeuze. Lijkt ingewikkeld doordat zinnen niet helemaal soepel lopen. Soms andere woorden gebruikt dan je zou verwachten als je normaal Nederlands schrijft.
108	Control	Het is allemaal te begrijpen, ook al heb ik het idee dat iemand die native speaker is van het Nederlands toch iets andere taal zou gebruiken. Misschien dat sommige dingen letterlijk zijn vertaald want sommige dingen klinken voor mij raar in deze tekst, terwijl ze in een andere taal wel normaal zouden klinken misschien.
109	Control	Woordkeuze is wat ongewoon, sommige zinnen hebben een kloppende, maar niet echt doorsnee structuur. De tekst las hierdoor niet helemaal lekker.
110	Control	Er zitten soms zinnen tussen die meer lijken op spreektaal dan op geschreven taal. Ook zitten er zinnen tussen die met te dure woorden geschreven zijn.
111	Control	De tekst is 'vlot' geschreven en is op een bepaalde manier ook wel grappig geschreven. De woordkeuze is soms wel apart of ver gezocht, zoals bijv. 'corporaties' ipv gewoon 'bedrijven'. Soms lijken zinnen of formuleringen niet helemaal te kloppen; het 'voelt' dan niet heel natuurlijk, maar het is volgens mij ook niet echt fout.
112	Control	In het begin was de tekst moeilijk te lezen of begrijpen, dat kwam denk ik door het gebruik van lange zinnen. Ook viel het mij op dat bij citaten de aanhalingstekens ontbraken of onjuist gebruikt waren. Ook werden er erg veel persoonlijk voornaamwoorden (mij, zij) gebruikt wat de tekst lastiger te lezen maakte.
113	Control	- rare, korte alinea's (wordt raar afgebroken); - vreemde zinnen; - rare/lastige woordkeuze
114	Control	Qua spelling is de tekst prima, maar qua woordkeuze en manier van formuleren zitten er veel fouten in. Veel woorden lijken een letterlijke vertaling uit het woordenboek en passen niet in de context van de tekst. Ook staan er interpunctie en stijlfouten in.
115	Control	Af en toe wat geforceerde zinnen waar de oorspronkelijke tekst 'doorheen schemert'. Of waaraan de lezer aanvoelt dat het geen xx heel Nederlandse zin is. Verder leest de tekst niet altijd even vloeiend. Er wordt opvallend veel 'zij' gebruikt (waar 'ze' vaak beter op zijn plaats was geweest) en de dialogen doen soms wat onnatuurlijk aan ("aangenaam jou weer te zien", "jij rustte bij het zitten gaan vol op je polsen" etc.)
116	Control	De verhaallijn is in het begin moeilijk op te pakken (relatief minder vlot dan in een goed lopend geschreven verhaal). Dit komt, denk ik, voornamelijk doordat zinnen niet altijd logisch opgebouwd zijn en ook niet altijd logisch/vloeiend aansluiten op de vorige zin. Sommige woorden komen in eerste instantie 'een beetje uit de lucht vallen', je moet ze dan nog een keer lezen om de inhoud te plaatsen. Woordkeuze varieert per zin. Soms 'gevorderde' woordenschat in een verder kromme zin -> dit bemoeilijkt ook het lezen. Teveel variatie in stijl?

D Picture description task

All pictures are originally from the Wasgij jigsaw puzzle range and were presented to the participants in full colour in A4-size.

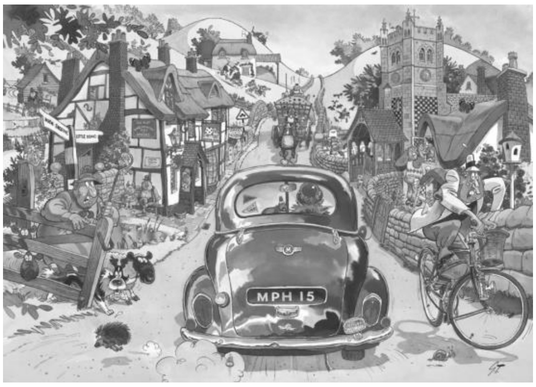


Figure 1. Picture description task, picture 1

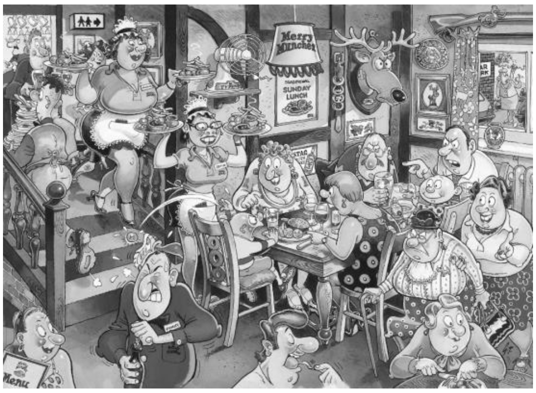


Figure 2. Picture description task, picture 2

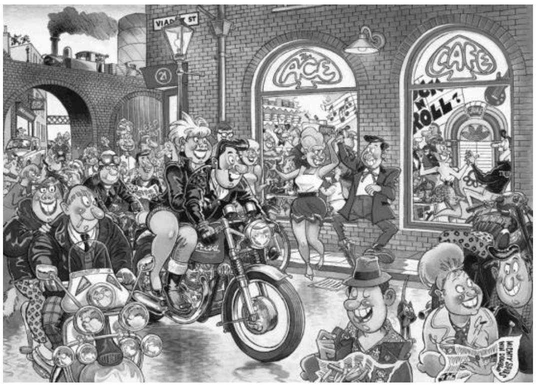


Figure 3. Picture description task, picture 3

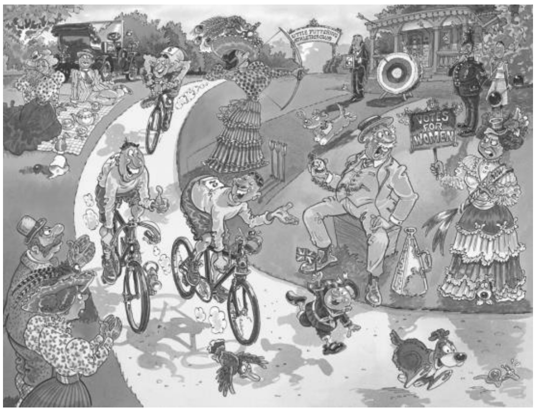


Figure 4. Picture description task, picture 4

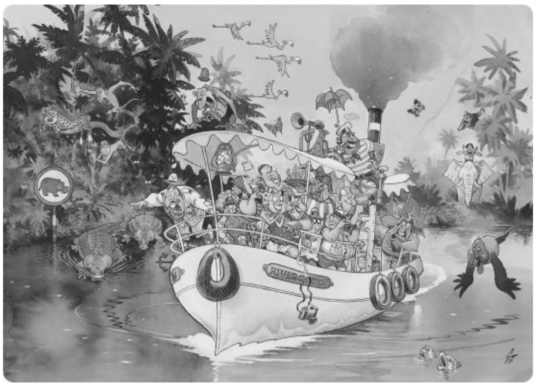


Figure 5. Picture description task, picture 5

E Parameter values of Zipf's law per transcript

Table E1. Parameter values for Zipf's law per transcript

		ZM-*α*			ZM-*β*			ZM-R2			Z-*β*			Z-R2		
	Sess.	**2000**	**1500**	**1278**	**2000**	**1500**	**1278**	**2000**	**1500**	**1278**	**2000**	**1500**	**1278**	**2000**	**1500**	**1278**
A02	1	1,13	1,093	1,066	4,714	3,93	3,081	0,992	0,976	0,977	1,757	1,717	1,778	0,988	0,986	0,99
A02	2	0,992	0,991	0,964	1,803	2,074	1,764	0,977	0,989	0,985	1,748	1,79	1,755	0,995	0,995	0,996
A02	3	1,06	1,001	0,964	2,061	1,195	0,745	0,977	0,981	0,985	1,728	1,875	1,904	0,99	0,991	0,988
A03	1		1,221	1,197		5,534	5,128		0,984	0,983		1,583	1,594		0,996	0,995
A03	2		1,124	1,082		4,465	4,028		0,975	0,974		1,753	1,791		0,993	0,99
A03	3	1,202	1,13	1,084	5,955	4,887	4,05	0,984	0,992	0,983	1,653	1,689	1,71	0,993	0,995	0,996
A04	1	1,046	0,967	0,935	2,254	1,168	0,838	0,981	0,983	0,98	1,934	1,972	1,996	0,993	0,992	0,995
A04	2		0,925	0,911		1,216	0,866		0,967	0,964		2,015	1,989		0,992	0,992
A04	3	0,981	0,951	0,951	1,69	1,33	1,311	0,991	0,975	0,981	1,99	1,996	2,004	0,995	0,994	0,992
A05	1		1,083	1,066		1,631	1,61		0,99	0,976		1,816	1,846		0,993	0,992
A05	2			1,089			2,397			0,99			1,786			0,988
A05	3		1,056	1,021		1,555	1,178		0,972	0,981		1,768	1,784		0,991	0,991
A06	1		0,983	0,961		1,253	1,132		0,96	0,956		1,71	1,704		0,991	0,991
A06	2		1,02	1,004		2,802	2,541		0,995	0,994		1,716	1,762		0,997	0,995
A06	3	1,107	1,036	0,99	4,254	3,322	2,663	0,995	0,992	0,994	1,708	1,768	1,806	0,992	0,991	0,994
A07	1	1,091	1,065	1,023	3,952	3,817	3,102	0,99	0,991	0,993	1,838	1,893	1,876	0,991	0,989	0,99
A07	2	1,12	1,086	1,078	3,466	3,046	3,21	0,983	0,983	0,983	1,88	1,942	1,928	0,991	0,989	0,989
A07	3	1,076	0,991	0,96	2,673	1,874	1,18	0,974	0,994	0,994	1,825	1,866	1,881	0,982	0,986	0,989
A08	1			1,04			2,067			0,985			1,683			0,995
A08	2			0,91			0,467			0,939			1,713			0,991
A08	3	1,058	1,045	1,021	2,077	2,551	2,51	0,964	0,973	0,979	1,688	1,713	1,719	0,995	0,996	0,996
C02		1,129	1,061	1,029	3,459	2,192	1,691	0,995	0,993	0,994	1,82	1,791	1,785	0,993	0,99	0,988
C03		1,171	1,112	1,11	5,955	4,828	4,592	0,977	0,978	0,978	1,779	1,789	1,797	0,984	0,992	0,992
C04		1,095	1,061	1,053	3,532	2,745	2,536	0,986	0,978	0,981	1,853	1,879	1,875	0,996	0,993	0,993
C05		1,076	1,038	1,023	3,109	2,476	2,327	0,973	0,975	0,987	1,792	1,757	1,767	0,994	0,993	0,994
C06		1,232	1,131	1,145	4,41	2,718	2,747	0,986	0,985	0,985	1,673	1,707	1,704	0,994	0,993	0,995
C07		1,019	0,953	0,909	2,407	1,424	0,955	0,984	0,976	0,955	1,945	1,941	2,005	0,994	0,995	0,996
C08		1,041	0,964	0,941	3,214	1,778	1,376	0,992	0,985	0,987	1,784	1,823	1,851	0,998	0,998	0,997

Bibliography

Avrutin, S. (2006). Weak syntax. In Y. Grodzinsky, & K. Amunds (Eds.), *Broca's region* (pp. 49-62) Oxford University Press.

Axer, H., Jantzen, J., Berks, G., Südfeld, D., & Keyserlingk, D. G. V. (2000). The aphasia database on the web: description of a model for problems of classification in medicine. In *Proc. ESIT*.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht/Boston/London: Kluwer Academic Publishers.

Baek, S. K., Bernhardsson, S., & Minnhagen, P. (2011). Zipf's law unzipped. *New Journal of Physics*, 13(4), 043004.

Baixeries, J., Elvevåg, B., & Ferrer i Cancho, R. (2013). The evolution of the exponent of Zipf's law in language ontogeny. *PLoS One, 8*(3), e53227.

Balasubrahmanyan, V. K., & Naranan, S. (2002). Algorithmic information, complexity and Zipf's law. *Glottometrics, 4*, 1-26.

Barabási, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(15 October), 509-512.

Baroni, M. (2008). Distributions in text. In A. Lüdelign, & M. Kytö (Eds.), *Corpus linguistics: An international handbook.* Berlin: Mouton de Gruyter.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G., Mazzoleni, M. (2004). Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. *Proceedings of LREC* 2004. http://sslmit.unibo.it/repubblica.

Baroni, M., & Ueyama, M. (2006). Building general- and special-purpose corpora by web crawling. *Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application*, 31-40.

Bastiaanse, R. (2011). The retrieval and inflection of verbs in the spontaneous speech of fluent aphasic speakers. *Journal of Neurolinguistics*, 24, 163–172.

Bastiaanse, R., Wieling, M., & Wolthuis, N. (2016). The role of frequency in the retrieval of nouns and verbs in aphasia. *Aphasiology*, 30(11), 1221-1239.

Bates, E. A., Friederici, A. D., Wulfeck, B. B., & Juarez, L. A. (1988). On the preservation of word order in aphasia: Cross-linguistic evidence. *Brain and Language*, 33, 323–364.

Belke, E., Brysbaert, M., Meyer, A.S., Ghyselinck, M. (2005). Age of acquisition effects in picture naming: Evidence for a lexical-semantic competition hypothesis. *Cognition*, *96*, B45-B54.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 57, 289–300.

Bentz, C., Alikaniotis, D., Samardžić, T., & Buttery, P. (2017). Variation in word frequency distributions: Definitions, measures and implications for a corpus-based language typology. *Journal of Quantitative Linguistics*, , 1-35.

Bentz, C., Kiela, D., Hill, F., & Buttery, P. (2014). Zipf's law and the grammar of languages: A quantitative study of old and modern English parallel text. *Corpus Linguistics and Linguistic Theory, 10*(2), 175-211.

Boxum, E., Van der Scheer, F. & Zwaga, M. (2010). *Analyse voor spontane taal; standaard in samenspraak met de VKL*. Vereniging voor Klinische Linguïstiek.

British National Corpus, version 3 (BNC XML Edition) (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition, 13*(7-8), 992-1011.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception & Performance, 42*(3), 441-458.

Brysbaert, M., Wijnendaele, I. van, & Deyne, S. de (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica, 104*, 215-226.

Budapest Sociolinguistic Interview (BUSZI), Hungarian Research Institute for Linguistics, Hungarian Academy of Sciences (RIL HAS). http://buszi.nytud.hu/.

Bybee, J., & Hopper, P. J. (Eds.) (2001). *Frequency and the emergence of linguistic structure*. Philadelphia: John Benjamins Publishing Company.

Caplan, D. (1987). *Neurolinguistics and linguistic aphasiology: An introduction*. Cambridge: Cambridge University Press.

Caramazza, A., Costa, A., Miozzo, M., & Bi, Y. (2001). The specific-word frequency effect: Implications for the representation of homophones in speech production. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*(6), 1430-1450.

Carroll, J. B., & White, M. N. (1973). Word frequency and age-of-acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology, 25,* 85-95.

CGN-Consortium. (2004). *Corpus Gesproken Nederlands*. Den Haag: Nederlandse Taalunie.

Chitashvili, R. J., & Baayen, R. H. (1993). Word frequency distributions. In G. Altmann, & L. Hřebíček (Eds.), *Quantitative text analysis* (pp. 54-135). Trier: WVT Wissenschatlicher Verlag.

Condon, E. U. (1928). Statistics of vocabulary. *Science*, (67), 300.

Corominas-Murtra, B., & Solé, R. V. (2010). Universality of Zipf's law. *Physical Review E, 82*, 011102-1-011102-9.

Corpus of Spoken Greek. Institute of Modern Greek Studies, Aristotle University of Thessaloniki, Greece. http://corpus-ins.lit.auth.gr/corpus/index.html.

Cuetos Vega, F., González Nosti, M., Martínez Jiménez, L., Mantiñán, N., & Olmedo, A. (2010). ¿Síndromes o síntomas en la evaluación de los pacientes afásicos? [Syndromes or symptoms in the assessment of aphasic patients?] *Psicothema, 22*(4), 715-719.

Dahui, W., Menghui, L., & Zengru, D. (2005). True reason for Zipf's law in language. *Physica A, 358*, 545-550.

Dell, G.S. (1986). A spreading activation theory of retrieval in language production. *Psychological review, 93*, 283-321.

Dell, G.S., & O'Seaghdha, P.G. (1992). Stages in lexical access in language production. *Cognition, 42*, 287-314.

De Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5), 292-306.

Egmond, M. van, Ewijk, L. van, & Avrutin, S. (2015). Zipf's law in non-fluent aphasia. *Journal Journal of Quantitative Linguistics, 22*(3), 233-249.

Ellis, A. W. (2006). Word finding in the damaged brain: Probing Marshall's caveat. *Cortex, 42*, 817-822.

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*(5), 1103-1123.

Estoup, J. B. (1916). *Gammes St´enographique*. Paris.

Ewijk, L. van (2013). *Word retrieval in acquired and developmental language disorders: A bit more on processing*. PhD dissertation, Utrecht University, LOT Dissertation Series 335.

Farmer, J. D., & Geanakoplos, J. (2008). Power laws in economics and elsewhere (tech. rep.). *Santa Fe Institute Tech Report,*

Fellbaum, C. (1998, ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Ferrer i Cancho, R. (2005a). The variation of Zipf's law in human language. *The European Physical Journal B, 44*, 249-257.

Ferrer i Cancho, R. (2005b). Zipf's law from a communicative phase transition. *The European Physical Journal B, 47*, 449-457.

Ferrer i Cancho, R. (2006). On the universality of Zipf's law for word frequencies. In P. Grzybek, & R. Köhler (Eds.), *Exact methods in the study of language and text. In honor of Gabriel Altmann* (pp. 131-140). Berlin: Gruyter.

Ferrer i Cancho, R. (2010). Information theory. In P. Colm Hogan (Ed.), *The Cambridge encyclopedia of the language sciences*. Cambridge: Cambridge University Press.

Ferrer i Cancho, R., & Elvevåg, B. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution. *PloS ONE, 5*(3), *e9411.*

Ferrer i Cancho, R., & Solé, R. V. (2000). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics, 8*(3), 165-173.

Ferrer i Cancho, R., & Solé, R. V. (2002). Zipf's law and random texts. *Advances in Complex Systems, 05*(01), 1-6.

Ferrer i Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America, 100*(3), 788-791.

Feyereisen, P., Van der Borght, F., & Seron, X. (1988). The operativity effect in naming: A re-analysis. *Neuropsychologia, 26,* 401-425.

Fromkin, V.A. (1971). The non-anomalous nature of anomalous utterances. *Language, 47,* 27-52.

FTD talk (www.ftdtalk.org), Factsheet 5 (version 1). Accessed on February 18, 2018.

Furusawa, C., & Kaneko, K. (2003). Zipf's law in gene expression. *Physical review letters*, 90(8), 088102.

Garrett, M.F. (1976). Syntactic processes in sentence production. *New approaches to language mechanisms*, 30, 231-256.

Garret, M.F. (1975). The analysis of sentence production. In: G.H. Bower (Ed.), *The psychology of learning and motivation. Vol. 9* (pp. 133-177). New York: Academic Press.

Gelbukh, A., & Sidorov, G. (2001). Zipf and Heaps laws' coefficients depend on language. *Proceeding of Conference on Intelligent Text Processing and Computational Linguistics (CICLing'2001), 2004.* pp. 332-335.

Gerhand, S., & Barry, C. (2000). When does a deep dyslexic make a semantic error? The roles of age-of-acquisition, concreteness, and frequency. *Brain and Language, 74,* 26-47.

Ghyselinck, M., Custers, R., & Brysbaert, M. (2004). The effect of age of acquisition in visual word processing: Further evidence for the semantic hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 550-554.

Gilhooly, K. J., & Gilhooly, M. L. M. (1980). The validity of age-of-acquisition ratings. *British Journal of Psychology*, 71, 105-110.

Goodglass, H. (1993). *Understanding aphasia.* New York: Academic Press.

Goodglass, H., Fodor, I. G., & Schulhoff, C. (1967). Prosodic factors in grammar: Evidence from aphasia. *Journal of Speech, Language, and Hearing Research*, 10, 5–20.

Goodglass, H., & Wingfield, A. (1997). Chapter 1 - Word-finding deficits in aphasia: Brain-behavior relations and clinical symptomatology. In: H. Goodglass, & A. Wingfield (Eds.), *Anomia* (pp. 3-27). San Diego: Academic Press.

Goutsos, D., Potagas, C., Kasselimis, D., Varkanitsa, M., & Evdokimidis, I. (Eds.). (2011). *Studying paraphasias in the Corpus of Greek Aphasic Speech.* Athens: Synapses.

Grodzinsky, Y. (1995). Trace deletion, θ-roles, and cognitive strategies. *Brain and language,* 51(3), 469-497.

Guiraud, P. (1971). The semic matrices of meaning. In J. Kristeva, J. Rey-Debove & D. J. Umiker (Eds.), *Essays in Semiotics/Essais de sémiotique* (pp. 150-159). The Hague/Paris: Mouton.

Guiraud, P. (1968). The semic matrices of meaning. *Social Science Information, 7*(2), 131-139.

Ha, L. Q., Stewart, D. W., Hanna, P., & Smith, F. J. (2006). Zipf and type-token rules for the English, Spanish, Irish and Latin languages. *Web Journal of Formal, Computational & Cognitive Linguistics,* (8)

Harley, T.A. (1984). A critique of top-down independent levels models of speech production: Evidence from non-plan internal speech errors. *Cognitive Science*, *8*, 191-219.

Harley, T.A., & MacAndrew, S.B.G. (1992). Modelling paraphasias in normal and aphasic speech. *Proceedings of the 14th annual conference on the Cognitive Science Society* (pp. 378-383). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Hatzigeorgiu, N., Mikros, G., & Carayannis, G. (2001). Word length, word frequencies and Zipf's law in the Greek language. *Journal of Quantitative Linguistics, 8*(3), 175-185.

Hernández-Fernández, A., & Diéguez-Vide, F. (2013). Zipf's law and the detection of the verbal evolution in Alzheimer's disease. *Anuario De Psicología/The UB Journal of Psychology, 43*(1), 67-82.

Howes, D. (1964). Application of the word-frequency concept to aphasia. In A. V. S. D. Reuck, & M. O'Connor (Eds.), *Disorders of Language*. Ciba Foundation Symposium (pp. 47–78). London: J. & A. Churchill, LTD.

Howes, D. (1968). Zipf's law and Miller's random-monkey model. *The American Journal of Psychology, 81*(2), 269-272.

Howes, D., & Geschwind, N. (1964). Quantitative studies of aphasic language. *Research Publications – Association for Research in Nervous and Mental Disease,* 42, 229–244.

Huang, W. (2014). Word frequency distribution in genres of modern Chinese. *QUALICO 2014 Book of Abstracts,* Olomouc, Czech Republic. pp. 63-64.

Izsák, F. (2006). Maximum likelihood estimation for constrained parameters of multinomial distributions—Application to Zipf–Mandelbrot models. *Computational Statistics & Data Analysis, 51*(3), 1575-1583.

Jäger, G., & van Rooij, R. (2007). Language structure: Psychological and social constraints. *Synthese, 159*(1), 99-130.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*(4), 824-843.

Jescheniak, J.D., Meyer, A.S., & Levelt, W.J.M. (2003). Specific-word frequency is not all that counts in speech production: Comments on Caramazza, Costa et al. (2001) and new experimental data. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 432-438.

Johnson, R.A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition*, *13*, 789-845.

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin, 131*(5), 684-712.

Juhasz, B.J., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1312-1318.

Keijzer, M. (2007). *Last in first out? An investigation of the regression hypothesis in Dutch emigrants in Anglophone Canada*. PhD dissertation, Vrije Universiteit Amsterdam, LOT Dissertation Series 163.

Kello, C. T., Brown, G. D. A., Ferrer i Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., et al. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences, 14*(5), 223-232.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new frequency measure for dutch words based on film subtitles. *Behavior Research Methods, 42*(3), 643-650.

Kittredge, A. K., Dell, G. S., Verkuilen, J., & Schwartz, M. F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology, 25*(4), 463-492.

Kostiç, A., Markovic, T., & Baucal, A. (2003). Inflectional morphology and word meaning: Orthogonal or co-implicative domains? In: R. H. Baayen & R.

Schreuder (Eds.), *Morphological structure in language processing* (pp. 1-44). Berlin: Mouton de Gruyter.

Krugman, P. (1996). Confronting the mystery of urban hierarchy. *Journal of the Japanese and International economies*, 10(4), 399-418.

Laine, M., & Martin, N. (2006). *Anomia. Theoretical and clinical aspects*. Hove and New York: Psychology Press, Taylor & Francis Group.

Laine, M., Tikkala., A., & Juhola, M. (1998). Modelling anomia by the discrete two-stage word production architecture. *Journal of Neurolinguistics*, *10*, 139-158.

Lees, R. B. (1959). Review: *Logique, language et théorie de l'information* by Léo Apostel; Benoit Mandelbrot; Albert Morf. *Language, 35*(2, Part 1), 271-303.

Lestrade, S. (2017). Unzipping Zipf's law. *PloS ONE, 12*(8), e0181987.

Levelt, W. J. M. (2013). *A history of psycholingusitics. The pre-chomkyan era.* Oxford University Press.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(1), 1-75.

Li, W. (2002). Zipf's law everywhere. *Glottometrics, 5*, 14-21.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *Information Theory, IEEE Transactions on, 38*(6), 1842-1845.

Li, W., Miramontes, P., & Cocho, G. (2010). Fitting ranked linguistic data with two-parameter functions. *Entropy, 12*, 1743-1764.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, 25, 1286–1307.

MacWhinney, B., & Osmán-Sági, J. (1991). Inflectional marking in Hungarian aphasics. *Brain and Language*, 41, 165–183.

Mandelbrot, B. B. (1953). An informational theory of the statistical structure of language. In: W. Jackson (Ed.), *Communication theory. papers read at a symposium on "applications of communication theory" held at the institution of electrical engineers, London September 22nd-26th 1952* (pp. 486-502). London: Butterworths Scientific Publications.

Mandelbrot, B. B. (1954). On recurrent noise limiting coding. Paper presented at the *Proc. Symp. on Inf. Networks, Polytechn. Inst. of Brooklyn*, pp. 205-221.

Manin, D. Y. (2009). Mandelbrot's model for Zipf's law: Can Mandelbrot's model explain Zipf's law for language? *Journal of Quantitative Linguistics, 16*(3), 274-285.

Manin, D. Y. (2008). Zipf's law and avoidance of excessive synonymy. *Cognitive Science, 32*(7), 1075-1098.

Marshall, J.C. (1977). Disorders in the expression of language. In: J. Morton and J.C. Marshall (eds.), *Psycholinguistics Series* (vol. 1). London: Elek Science, 1977.

McClelland, J.L., McNaughton, B.L., & O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the success and failures of connectionist models of learning and memory. *Psychological Review*, *88*, 375-407.

McNeil, M. R., Odell, K., & Tseng, C. H. (1991). Toward the integration of resource allocation into a general theory of aphasia. *Clinical aphasiology*, *20*, 21-39.

Meyer, P. (2002). Laws and theories in quantitative linguistics. *Glottometrics, 5*, 62-80.

Miller, G.A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), 39-41.

Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology, 70*(2), 311-314.

Miller, G. A., & Newman, E. B. (1958). Tests of a statistical explanation of the rank-frequency relation for words in written English. *The American Journal of Psychology, 71*(1), 209-218.

Miller, G. A., Newman, E. B., & Friedman, E. A. (1958). Length-frequency statistics for written English. *Information and Control, 1*(4), 370-389.

Monaghan, J., & Ellis, A.W. (2002). Age of acquisition and the completeness of phonological representations. *Reading and Writing*, *15*, 759-788.

Montemurro, M. A. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A, 300*, 567-578.

Montemurro, M. A., & Zanette, D. (2002). Entropic analysis of the role of words in literary texts. *Advances in Complex Systems, 05*(01), 7-17.

Montemurro, M. A., & Zanette, D. H. (2011). Universal entropy of word ordering across linguistic families. *PLoS One, 6*(5), e19875.

Moor, W. de, Ghyselinck, M., & Brysbaert, M. (2000). A validation study of the age-of-acquisition norms collected by Ghyselinck, De Moor, & Brysbaert. *Psychologica Belgica*, (2), 114.

Morton, J. (1970). A functional model for human memory. In: D.A. Norman (Ed.), *Models of human memory* (pp. 203-260). New York: Academic Press.

Morton, J., & Patterson, K. (1980). A new attempt at an interpretation or an old attempt at a new interpretation. In: M. Coltheart, K. Patterson, & J.C. Marshall (Eds.), *Deep dyslexia* (pp. 91-118). London: Routledge & Kegan Paul.

Murphy, L. (2015). *Likelihood: Tools for maximum likelihood estimation.* R package version 1.7.

Naranan, S., & Balasubrahmanyan, V. (1992). Information theoretic models in statistical linguistics—Part I: A model for word frequencies. *Current Science*, 63(5), 261-269.

Nazir, T.A., Decoppet, N., & Aghababian, V. (2003). On the origins of age-of-acquisition effects in the perception of printed words. *Developmental Science, 6*, 143-150.

Németh, G., & Zainkó, C. (2002). Multilingual statistical text analysis, Zipf ‘s law and Hungarian speech generation. *Acta Linguistica Hungarica*, 49, 385–405.

Neophytou, K., Egmond, M. van & Avrutin, S. (2017). Zipf's law in aphasia across languages: A comparison of English, Hungarian and Greek. *Journal of Quantitative Linguistics*, 24(2-3), pp. 178-196.

Newman, M. (2005). Power laws, pareto distributions and Zipf’s law. *Contemporary Physics, 46*(5), 323-351.

Nickels, L. (1995). Getting it right? Using aphasic naming errors to evaluate theoretical models of spoken word production. *Language and Cognitive Processes*, *10*(1), 13-45.

Nickels, L., & Howard, D. (1994). A frequent occurrence? Factors affecting the production of semantic errors in aphasic naming. *Cognitive Neuropsychology*, *11*, 289-320.

Nickels, L., & Howard, D. (1995). Aphasic naming: What matters? *Neuropsychologia, 33*, 1281-1303.

Nieuwenhuis, S. (2005). *Ik ben omringd door debielen en ik voel me goed.* Passage.

NIST/SEMATECH e-Handbook of Statistical Methods (accessed: April 11, 2018). www.itl.nist.gov/div898/handbook/prc/section4/prc471.htm.

Oettinger, A. G. (1954). The distribution of word length in technical Russian. *Mechanical Translation, 1*, 38-40.

Okuyama, K., Takayasu, M., & Takayasu, H. (1999). Zipf’s law in income distribution of companies. *Physica A: Statistical Mechanics and its Applications*, 269(1), 125-131.

Oxford English Dictionary (2017), Dictionary Facts. Oxford University Press. http://public.oed.com/history-of-the-oed/dictionary-facts/ (accessed: February 2018)

Piantadosi, S. T. (2014). Zipf’s word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review,* 1-19.

Piotrovskii, R. G., Pashkovskii, V. E., & Piotrovskii, V. R. (1994). Psychiatric linguistics and automatic text processing. *Nauchno-Tekhnicheskaya Informatsiya, Seriya 2, 28*(11), 21-25.

Piotrowski, R. G., & Spivak, D. L. (2007). Linguistic disorders and pathologies: Synergetic aspects. In P. Grzybek, & R. Köhler (Eds.), *Exact methods in the study of language and text. in honour of Gabriel Altman* (pp. 545-554). Berlin: Walter de Gruyter.

Plaut, D. (1996). Relearning after damage in connectionist networks: Toward a theory of rehabilitation. *Brain and Language, 52*, 25-82.

Plaut, D., & Shallice, T. (1993a). Perseverative and semantic influences on visual object naming errors in optic aphasia: A connectionist account. *Journal of Cognitive Neuroscience*, *5*, 89-117.

Plaut, D., & Shallice, T. (1993b). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377-500.

Popescu, I.-I., Altmann, G., & Köhler, R. (2010). Zipf's law - another view. *Quality and Quantity, 44*, 713-731.

R Core Team and contributors worldwide (2017), documentation concerning the R stats package 3.4.3. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html (accessed: January 2018).

Rapoport, A. (1982). Zipf's law re-visited. In H. Guiter, & M. V. Arapov (Eds.), *Studies on Zipf's law* (pp. 1-28). Bochum: Studienverlag Dr. N. Brockmeyer.

Rapp, B., & Coldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107*, 406-499.

Richie, R. (2016). Functionalism in the lexicon: Where is it, and how did it get there? *The Mental Lexicon, 11*(3), 429-466.

Richie, R., Kaufmann, S., & Tabor, W. (2014). An LSA-based method for estimating word meaning specificity: An application to an account of Zipf's law. Poster Presented at the *9th Annual Mental Lexicon Conference,* Niagara-on-the-Lake, Ontario, Canada.

Ridley, D. R. (1982). Zipf's law in transcribed speech. *Psychological Research, 44*(1), 97-103.

Ridley, D. R., & Gonzales, E. A. (1994). Zipf's law extended to small samples of adult speech. *Perceptual and Motor Skills, 79*, 153-154.

Roelofs, A. (1997). The weaver model of word-form encoding in speech production. *Cognition*, 64(3), 249-84.

Roelofs, A. (1996). Serial order in planning the production of successive morphemes of a word. *Journal of Memory and Language*, 35(6), 854-876.

Roelofs, A. (1993). Testing a non-decompositional theory of lemma retrieval in speaking: Retrieval of verbs. *Cognition*, 47(1), 59-87.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, 42(1-3), 107-42.

Roelofs, A., & Meyer, A. S. (1998). Metrical structure in planning the production of spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 922-939.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379-423, 623-656.

Simon, H. A. (1960). Some further notes on a class of skew distribution functions. *Information and Control*, 3, 80-88.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika, 42*(3-4), 425-440.

Smith, M. A., Cottrell, G. W., & Anderson, K. L. (2001). The early word catches the weights. *Advances in neural information processing systems*. Cambridge.

Sornette, D. (2006). Mechanisms for power laws. *Critical phenomena in natural sciences* (pp. 345-394). Berlin; Heidelberg: Springer.

Springer, L. (2008). Therapeutic approaches in aphasia rehabilitation. In B. Stemmer, & H. A. Whitaker (Eds.), *Handbook of the neuroscience of language* (pp. 397-406), Elsevier Science & Technology.

Stemberger, J.P. (1985). An interactive model of language production. In: A.W. Ellis (Ed.), *Progress in the psychology of language*, *vol. 1* (pp. 143-186). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41-78.

Tabak, W., Schreuder, R., & Baayen, R. H. (2010). Producing inflected verbs: A picture naming study. *The Mental Lexicon* 5(1), 22-46.

Tabak, W., Schreuder, R., & Baayen, R. H. (2006). *Nonderivational inflection.* Manuscript, Max Planck Institute for Psycholinguistics.

The British National Corpus, *version 3* (BNC XML Edition) (2007). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/.

Tripp, O., & Feitelson, D. (2007). *Zipf's law revisited*. School of Computer Science and Engineering, The Hebrew University of Jerusalem, Tech. Report: No. 115. Retrieved from http://researcher.ibm.com/researcher/files/us-otripp/tr07.pdf.

Turner, J. E., Valentine, T., & Ellis, A. W. (1998). Contrasting effects of age of acquisition and word frequency on auditory and visual lexical decision. *Memory & Cognition, 26*(6), 1282-1291.

Tuzzi, A., Popescu, I., & Altmann, G. (2009). Zipf's law in Italian texts. *Journal of Quantitative Linguistics, 16*(4), 354-367.

Tweedie, F.J. & Baayen, R.H. (1998). How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, *32*(5), 323-352.

Vogt, P. (2004). Minimum cost and the emergence of the Zipf-Mandelbrot law. *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*. Cambridge, MA: MIT Press.

Wernicke, C. (1874). The aphasia symptom complex: A psychological study on an anatomical basis. Reprinted in G. Eggert (1977). *Wernicke's works on aphasia: A sourcebook and review.* Berlin: Mouton de Gruyter.

Wyllys, R. E. (1981). Empirical and theoretical bases of Zipf's law. *Library Trends, 30*(1), 53-64.

Yang, C. (2013). Who's afraid of George Kingsley Zipf? *Significance, 10*(6), 29-34.

Zanette, D., & Montemurro, M. (2005). Dynamics of text generation with realistic Zipf's distribution. *Journal of Quantitative Linguistics, 12*(1), 29-40.

Zevin, J.D. & Seidenberg, M.S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language, 47,* 1-29.

Zipf, G. K. (1949). *Human behavior and the principle of least effort. An introduction to human ecology*. New York and London: Hafner Publishing company.

Zurif, E., Swinney, D., Prather, P., Solomon, J., & Bushell, C. (1993). An on-line analysis of syntactic processing in Broca's and Wernicke's aphasia. *Brain and Language,* 45, 448–464.

Samenvatting in het Nederlands

In dit proefschrift neem ik de wet van Zipf onder de loep in verschillende bijzondere gevallen. De wet van Zipf gaat over de volgende wetmatigheid. Neem een willekeurige geschreven tekst van enige lengte (bijvoorbeeld een roman). Tel van elk woord hoe vaak het voorkomt, en noteer dit. Sorteer deze lijst, van meest frequent naar minst frequent, en geef de woorden rangnummers: het meest frequente woord krijgt rang 1, het op één na meest frequente woord krijgt rang 2, en zo verder totdat alle woorden een nummer hebben. Er wordt nu een patroon zichtbaar: het meest frequente woord is ongeveer twee keer zo veelvoorkomend als het op één na meest frequente woord, en dit woord is weer ongeveer twee keer zo veelvoorkomend als het op twee na meest frequente woord. Ongeveer de helft van alle woorden komt maar één keer voor in de tekst. Dit patroon wordt beter zichtbaar in een grafiek met logaritmische schaalverdelingen, met rangnummers op de x-as en woordfrequenties op de y-as. Nu is te zien dat de frequentiewaardes log-lineair verdeeld zijn: ze vallen (bij benadering) op één rechte lijn (zie figuur 2.2, p. 21, voor een voorbeeld). Dit patroon wordt niet alleen gevonden voor Nederlandstalige teksten, maar voor elke tekst in elke natuurlijke taal die tot nu toe onderzocht werd. Dit verschijnsel staat bekend als de *Wet van Zipf*.

De wet van Zipf is veelvuldig onderzocht in grote corpora en lange teksten. De vraag die ik mezelf stelde was: Gaat deze wetmatigheid ook op wanneer de taalproductie gestoord is, zoals in de spraak van mensen met de verworven taalstoornis afasie? Echter was er tot dusverre vrijwel niets bekend over de wet van Zipf in kleine, gesproken teksten. Aan deze vraag gaan daarom andere vragen vooraf. Bestaan er structurele verschillen tussen gesproken en geschreven teksten? Wat is de invloed van het medium van de tekst? Wat is de minimale tekstlengte waarin de wet van Zipf onderzocht kan worden? Is het mogelijk om een tekst te construeren waarin de wet van Zipf *niet* van toepassing is, en hoe reageren lezers hierop? Dit zijn de vragen waarop ik me in dit proefschrift richt. Mijn onderzoeksvragen zijn uitgewerkt in hoofdstuk 1. In hoofdstuk 2 bespreek ik de theorie van de wet van Zipf. In hoofdstuk 3 onderzoek ik de werking van de wet van Zipf in kleine gesproken en geschreven teksten. Daarna richt ik me in hoofdstuk 4 en 5 op de wet van Zipf in afasie: in hoofdstuk 4 doe ik dat in lange spraakfragmenten in het Nederlands, in hoofdstuk 5 richt ik me op korte spraakfragmenten in het Engels, Grieks en Hongaars. In hoofdstuk 6 reflecteer ik op mijn bevindingen.

In hoofdstuk 2 bespreek ik de twee theoretische pijlers onder dit onderzoek: spraakproductie in mensen met en zonder afasie en de verschillende theorieën voor het bestaan van de Wet van Zipf.
Voor spraakproductie ga ik uit van Levelt's model. Volgens dit model worden woorden geproduceerd in vijf stappen in twee domeinen. Conceptuele voorbereiding en lexicale selectie vinden plaats in het conceptuele/syntactische domein; morfologische codering en syllabificatie, fonetische codering en articulatie vinden plaats in het fonologische/articulatorische domein. Deze processen vinden plaats door middel van activatie van opeenvolgende (sets van) knopen in een netwerk.

Twee factoren die van grote invloed zijn op het gemak waarmee woorden geproduceerd worden zijn hun frequentie in de taal, en de leeftijd waarop ze verworven zijn. Woorden die vaker voorkomen in een taal worden sneller, gemakkelijker en/of beter verwerkt dan woorden die zeldzamer zijn. Dit effect staat bekend als het frequentie-effect. Het is onbekend waar in het proces van spraakproductie dit effect precies ontstaat, maar het lijkt het geval te zijn dat hier meerdere niveaus bij betrokken zijn, met name op het fonologische niveau.
Daarnaast worden woorden die als kind eerder geleerd zijn sneller, gemakkelijker en/of beter verwerkt dan woorden die later geleerd zijn. Dit staat bekend als het Age of Acquisition (AoA) effect. Over dit effect bestaat zo mogelijk nog meer onduidelijkheid. Waarschijnlijk is neurologische plasticiteit ten minste ten dele verantwoordelijk voor dit effect: doordat de hersenen van jonge kinderen een grotere plasticiteit hebben bepalen 'vroege' woorden de uiteindelijke structuur van het netwerk als geheel wanneer dit volgroeid is.

Hersenletsel kan de taalstoornis afasie tot gevolg hebben. Deze aandoening omvat een breed scala aan symptomen, die afhankelijk van de exacte locus van de hersenbeschadiging voor elk persoon anders kunnen zijn. Afasievormen kunnen onderverdeeld worden in vloeiende en niet vloeiende afasie. Iemand met vloeiende afasie produceert grammaticaal kloppende zinnen, maar uit vaak de verkeerde woorden en gebruikt neologismen. Hierdoor is het voor een ander vaak moeilijk te begrijpen. Iemand met niet-vloeiende afasie spreekt vaak in wat 'telegraafstijl' genoemd wordt: de spraak klinkt hakkelend en de spreker heeft last van woordvindingsproblemen: de spreker weet wat hij of zij wil zeggen, maar komt niet op het woord. Woordproductie is dus gestoord. Men gaat ervan uit dat dit komt door moeilijkheden bij het ophalen van woorden uit het geheugen; de woorden zijn dus niet vergeten en het netwerk van woorden in het geheugen is intact.

De tweede pijler onder dit onderzoek is de wet van Zipf. De versie van deze wet die hier gebruikt wordt is die na aanpassing door Mandelbrot, bekend als de Zipf-Mandelbrot law:

1) $f(w) \approx \frac{C}{(r(w)+\beta)^{\alpha}}$

In deze formule wordt de woordfrequentie *f* van woord *w* bepaald door een constante *C* die met name afhankelijk is van de lengte van de tekst, de rang *r* en de parameters α en β. Parameter α (hier verder aangeduid als ZM-α) bepaalt de hellingshoek van de curve, en ligt meestal rond ZM-$\alpha \approx 1$. Parameter β (hier verder aangeduid als ZM-β) bepaalt de mate waarin de curve voor de eerste paar rangen afwijkt van de gegeven hellingshoek: de woordfrequenties zijn voor deze rangen doorgaans lager dan gegeven door de lijn met hellingshoek ZM-α. De waarde van ZM-β is afhankelijk van de tekst, maar is zelden groter dan 10. De wet van Zipf is een specifiek voorbeeld van een machtsverband.

Daarnaast bestaat er ook een andere versie van de wet van Zipf, hier Zipf's β-law genoemd:

2) $n_f \approx C \cdot f^{-\beta}$

Deze formulering betreft het aantal woorden *n* met frequentie *f*, dat bepaald wordt door de constante *C* die afhankelijk is van tekstgrootte, en parameter β (hier verder aangeduid als Z-β) die meestal rond $\beta \approx 2$ ligt.

De verschillende hypotheses voor het bestaan van de wet van Zipf kunnen ingedeeld worden in vijf groepen. Klassieke verklaringen gaan uit van het principe van de minste moeite (*Principle of Least Effort*). Dit was G.K. Zipf's eigen verklaring voor de wet die zijn naam draagt. Het idee is dat sprekers de minste moeite hoeven doen als ze één woord gebruiken voor alles wat ze willen zeggen. Luisteraars hebben juist liever voor elke betekenis een uniek woord. De spanning tussen deze belangen resulteert in de wet van Zipf.

Verklaringen gebaseerd op periodieke pauzes (*Intermittent Silence*) gebruiken een argument in lijn met de aap die op een toetsenbord slaat. Deze denkbeeldige aap gebruikt de spatiebalk met waarschijnlijkheid *p*(*) en alle ander toetsen met waarschijnlijkheid p(L) = 1 – *p*(*). Hij gebruikt nooit twee keer de spatiebalk na elkaar. Het resultaat zal dan bestaan uit 'woorden' van *i* letters op een rij. De waarschijnlijkheid van een woord met *i* letters neemt exponentieel af als *i* toeneemt. Met andere woorden: korte woorden komen veel vaker voor dan lange woorden. Op

logaritmische assen volgt deze verdeling een machtsverband, een rechte lijn.
Meer recente verklaringen werken op basis van *preferential attachment* in de onderliggende mentale netwerken. Vroeg geleerde woorden bepalen de structuur. Doordat nieuwe woorden aanhechten in de bestaande structuur verwerven de vroege woorden de meeste connecties en functioneren ze als ‘hubs’ in deze netwerken. Later geleerde woorden komen in de periferie van de netwerken terecht en hebben maar weinig verbindingen met andere woorden. Dit proces resulteert in een netwerk waarin enkele woorden heel veel connecties hebben, terwijl de meesten maar 1 of 2 connecties hebben. Het gebruik van een dergelijk onderliggend netwerk resulteert in teksten waarin enkele woorden hoogfrequent zijn, terwijl de meesten zeer laagfrequent zijn – met andere woorden, de wet van Zipf.
Andere recente verklaringen bouwen voort op Zipf’s Principle of Least effort op een manier die in zijn tijd nog niet mogelijk was. Dit zijn computationele modellen op basis van optimalisatie van de beschikbare en benodigde middelen.
Tot slot is er een gevarieerde groep aan semantische hypotheses, waarin de betekenis van de woordenverantwoordelijk gehouden wordt voor de typische woordfrequenties in teksten.
Ik kom tot de conclusie dat de hypotheses op basis van preferential attachment in mentale netwerken het meest plausibel zijn. Deze verklaringen zijn psychologisch en neurologisch het meest aannemelijk, en de beschreven processen kunnen niet alleen de wet van Zipf in taal verklaren, maar ook gelijksoortige fenomenen op andere gebieden.

In hoofdstuk 3 onderzoek ik de werking van de wet van Zipf in kleine, gesproken teksten. De verdeling in geschreven of gesproken teksten is geen dichotomie maar een continuüm: gesproken teksten kunnen sterk doordacht zijn (denk aan voordrachten); geschreven teksten kunnen spontaan en zonder revisie gepubliceerd zijn (denk aan blogs). Ik heb daarom teksten geanalyseerd op een continuüm van sterk gereviseerde geschreven teksten naar volledig spontane gesproken teksten, bestaande uit zeven categorieën: spontane gesprekken tussen familie of vrienden, spontane radio- of televisiecommentaren, discussies op radio of televisie, preken en toespraken, blogs, nieuwsartikelen, en literatuur. Voor al deze teksten heb ik groeicurves opgesteld: ik heb de waardes van de parameters van de wet van Zipf berekend voor een groeiend aantal woorden, beginnend bij 100 en eindigend bij het einde van de tekst of bij 5000 woorden. Doel hiervan was vaststellen wanneer deze waardes stabiliseren, om zo te kunnen bepalen bij welk aantal woorden het zinvol is om de waardes van de parameters te vergelijken tussen verschillende teksten. Het blijkt dat de waardes van ZM-α zich voor alle teksten op dezelfde wijze ontwikkelen wanneer het aantal geanalyseerde woorden toeneemt: in alle gevallen volgen de groeicurves een logaritmisch traject (wat inhoudt dat de waarde van ZM-α eerst zeer snel toeneemt om daarna bij benadering lineair door te groeien). Voor ZM-β is geen patroon te ontdekken, wat waarschijnlijk komt doordat deze parameter berekend wordt op basis van een klein aantal waardes (de frequenties van de hoogste rangen). De resultaten voor Z-β waren onverwachts: de groeicurves leken hetzelfde traject te volgen: na enige fluctuatie blijven de waardes vrijwel gelijk. Toch bleken ze niet met eenzelfde formule te beschrijven. Het gebrek aan fluctuatie voor langere teksten wijst erop dat het ondanks het ontbreken van een uniforme formule om de groei te beschrijven wel zinvol is om de waarde van de parameter *tussen* teksten te vergelijken. Uitgaande van de groeicurves voor ZM-α en Z-β stel ik vast dat teksten van ten minste 1500 woorden nodig zijn om de wet van Zipf te kunnen bestuderen.

Vervolgens heb ik de waardes van de parameters vergeleken tussen de teksten voor de eerste 1000, 1500 en 2000 woorden van elke tekst (voor zover beschikbaar, sommige teksten waren korter), om zo te onderzoeken of er systematische verschillen bestaan tussen gesproken en geschreven teksten. Dit blijkt zo te zijn: ZM-α blijkt hoger te zijn voor gesproken teksten dan voor geschreven teksten; voor Z-β geldt het tegenovergestelde. Dit verschil reflecteert een grotere woordenschat in geschreven dan in gesproken taal. Interessant genoeg geldt dit verschil ook voor sterk gereviseerde gesproken teksten, zoals voordrachten en preken. De samensteller van dit soort teksten houdt dus blijkbaar rekening met het feit dat de tekst uitgesproken zal worden. Blogs bevinden zich op een grenspositie.

In deel II van hoofdstuk 3 richt ik me op een heel andere experimentele benadering van de wet van Zipf. Tot nu toe werd deze wetmatigheid vastgesteld in elke tekst waarin deze onderzocht werd. Dit betekent dat er, voor zover tot nu toe bekend, geen natuurlijk voorkomende teksten zijn waarin de wet van Zipf *niet* voorkomt. Ik construeerde een dergelijke tekst om uit te vinden hoe lezers hierop reageren. Dit heb ik gedaan door eerst kunstmatig een frequentieverdeling op te stellen waar de aangepaste tekst vervolgens aan moest voldoen. Het blijkt dat Zipf's β-law ernstig verstoord kan worden, terwijl Zipf-Mandelbrots law maar in zeer beperkte mate verstoord is. Zipf's β-law is dus gevoeliger voor verstoringen dan Zipf-Mandelbrots law.

De tekst heb ik daarna aan de opgestelde frequentieverdeling laten voldoen door laagfrequente woorden te vervangen door passende hoogfrequente woorden en andersom. Om te voorkomen dat ongrammaticaliteit de resultaten zou beïnvloeden heb ik dit alleen voor inhoudswoorden (zelfstandige naamwoorden, werkwoorden, bijvoeglijke naamwoorden en bijwoorden) gedaan. Als gevolg hiervan was de wet van Zipf verstoord voor inhoudswoorden, maar niet voor functiewoorden. Deze tekst heb ik vervolgens voorgelegd aan een groep lezers. Zij moesten hierbij een vragenlijst invullen, gericht op een breed spectrum aan tekstkarakteristieken. De aangepaste tekst heb ik vergeleken met het origineel en een controletekst waarin andersoortige aanpassingen gedaan zijn (woorden zijn vervangen door hun synoniemen, en onbeklemtoonde persoonlijk voornaamwoorden [bijv. ‘je’] zijn vervangen door de beklemtoonde versies [bijv. ‘jij’]). Uit de resultaten bleek dat de veranderingen opgemerkt werden: de woordkeus op het gebied van werkwoorden, zelfstandige naamwoorden en bijvoeglijke en bijwoorden in de experimentele tekst werd beoordeeld als minder gevarieerd dan in de twee controleteksten. Het enige punt waarop de tekst zonder de wet van Zipf lager beoordeeld werd dan de controletekst was de vraag hoe poëtisch de tekst was. Mogelijk is dit te wijten aan de beperktere woordkeus en minder laagfrequente woorden, of het komt doordat het leesritme van de tekst verstoord is. Mogelijk was de tekst niet ‘verstoord’ genoeg om sterkere resultaten te bewerkstelligen.

Dit was een eerste, verkennende studie naar dit onderwerp. Een andere aanpak bij het verstoren van de wet van Zipf in een tekst (zoals bijvoorbeeld het zelf schrijven van een tekst in plaats van een bestaande tekst aanpassen) heeft mogelijk meer effect.

Met deze achtergronden kon ik de wet van Zipf onderzoeken in spraakfragmenten van mensen met niet-vloeiende vormen van afasie. Dit onderzoek is gerapporteerd in hoofdstuk 4. Als de wet van Zipf zijn origine vindt in het mentale netwerk, en dit netwerk is zoals hierboven gesteld intact in mensen met afasie, dan zou de wet van Zipf ook onverminderd moeten gelden voor de spraak van mensen met afasie. Naast de wet van Zipf heb ik hun taal ook geanalyseerd op de distributie van woordfrequenties in Nederlands in het algemeen (SUBTLEX-frequentie) en op Age of Acquisition, omdat die, zoals besproken in hoofdstuk 2, een rol spelen in de organisatie van het mentale lexicon.

De deelnemers waren vier mensen milde vormen van afasie, drie mensen met meer ernstige vormen van afasie en zeven gematchte controledeelnemers. Ik heb ca. 45 minuten spraak verzameld per persoon, aan de hand van een afbeeldingsbeschrijvingstaak, een filmbeschrijvingstaak en een vrij gesprek. Voor elk gesprek zijn de parameters van de wet van Zipf berekend. Voor die woorden waarvoor dit beschikbaar was zijn AoA-normen en SUBTLEX-frequentiewaardes geanalyseerd (via the Dutch Lexicon Project 2, Brysbaert, Mandera & Keuleers, 2016). Uit de analyses blijkt dat de wet van Zipf in gelijke mate opgaat voor spraak van beide groepen. Er is geen verschil wat betreft de waardes van de parameters. Dit is mogelijk te wijten aan de milde vorm van afasie van veel deelnemers: de waardes voor ZM-α voor de mensen met meer ernstige afasie zijn hoger dan die voor de deelnemers met mildere afasie. Voor Z-β zijn de resultaten gespiegeld. Dit verschil suggereert een kleinere actieve woordenschat in de gesprekken van mensen met (meer ernstige) afasie. De algemene vorm van de distributie is echter ongewijzigd: alle gesprekken volgen de wet van Zipf. Dit suggereert dat mensen met en zonder afasie woorden op dezelfde manier in hun geheugen opslaan en hieruit ophalen voor gebruik.

Voor AoA vind ik dat de gemiddelde verwervingsleeftijd van de gebruikte woorden lager ligt bij mensen met afasie. Dit resultaat is in lijn met de hypothese dat eerder verworven woorden dieper ingebed liggen in de mentale netwerken, waardoor ze gemakkelijker te gebruiken zijn en beter bestand zijn tegen beschadigingen. Voor SUBTLEX-frequentie vind ik dit verschil niet, maar vind ik een verschil tussen testsessies. Dit duidt erop dat het hier inderdaad om twee verschillende effecten gaat.

In hoofdstuk 5 verruim ik mijn blik en richt ik me op de wet van Zipf in afatische spraak in het Engels, Grieks en Hongaars. Tevens includeer ik nu spraak van mensen met vloeiende afasievormen. Voor het Engels was ook een controlegroep aanwezig.

De onderzochte spraakfragmenten komen uit het AphasiaBank Corpus (MacWhinney, Fromm, Forbes, & Holland, 2011). Dit corpus bevat helaas alleen korte samples, veel korter dan de eerder vastgestelde minimaal 1500 woorden die nodig zijn voor een juiste vergelijking van de parameters van de wet van Zipf. De geanalyseerde samples waren in alle gevallen 200 woorden lang. Om deze waardevolle dataset niet ongebruikt te hoeven laten stel ik een methode voor om de waardes van de wet van Zipf te valideren voor kleinere tekstfragmenten, om zo de gevonden waardes te kunnen interpreteren. Dit doe ik door middel van taalspecifieke baseline corpora. Deze corpora bevatten grote hoeveelheden spontane spraak van gezonde sprekers. Uit deze corpora heb ik zoveel mogelijk fragmenten van 200 woorden geanalyseerd, om zo de variantie binnen normale spraakfragmenten vast te kunnen stellen. Dit stelde mij in staat om de resultaten van de AphasiaBank fragmenten te duiden. Ik gebruik alleen Zipf-Mandelbrots law voor deze vergelijkingen, Zipf's β-law is ongeschikt voor korte teksten.

Op deze manier kan ik concluderen dat alle samples de wet van Zipf volgen, in lijn met mijn bevindingen voor het Nederlands. Maar de waardes van de parameters zijn in alle gevallen significant anders dan die van de baseline corpora, met uitzondering van de Engelse controle-deelnemers: ZM-α is hoger in spraakfragmenten van mensen met afasie dan in de fragmenten uit de baseline corpora. Dit verschil wijst op een kleiner vocabulaire in de spraaksamples van de mensen met afasie. Het verschil tussen groepen is meer uitgesproken dan dat tussen de Nederlandse groepen, mogelijk doordat de AphasiaBank deelnemers ernstigere vormen van afasie hadden. Opvallend is dat ik geen verschil vind tussen de spraak van mensen met vloeiende en niet vloeiende afasie. Kwalitatieve analyses van de spraakfragmenten zijn nodig om dit gebrek aan verschil juist te kunnen interpreteren.

De waardes van de parameters van de wet van Zipf zijn anders voor de drie verschillende talen. ZM-α is hoger voor Engels dan voor Grieks, en neigt hoger te zijn voor Engels dan voor Hongaars (p = 0,06). Tussen Grieks en Hongaars vind ik geen verschil. Dit effect is naar alle waarschijnlijkheid te wijten aan het verschil in morfologische verscheidenheid: de talen met lagere waardes voor ZM-α kennen meer mogelijkheden tot verbuiging en vervoeging van woorden, wat leidt tot het gebruik van meer verschillende woorden.

In hoofdstuk 6 reflecteer ik op de voorgaande hoofdstukken en beantwoord ik de onderzoeksvragen die ik stelde in hoofdstuk 1. Dit onderzoek werpt licht op de werking van de wet van Zipf in ‘randgevallen’: korte teksten, gesproken teksten en spraak van mensen met afasie. Ondanks de wijdverbreide acceptatie van de universaliteit van de wet van Zipf waren dit soort teksten tot nu toe vrijwel niet onderzocht. Ik heb in deze studie aangetoond dat de wet van Zipf ook voor deze gevallen geldt, maar niet met de waardes ZM-α = 1 en Z-β = 2 die vaak als universeel geldend worden aangenomen. Ik toon aan dat de waardes van deze parameters bepaald worden door tekstkarakteristieken zoals tekstlengte, genre, medium en vloeiendheid van de spreker. Ze zijn daarom voor elke tekst anders. Een tekstlengte van minimaal 1500 woorden is nodig om deze waardes onderling direct te kunnen vergelijken. De waardes voor kortere teksten kunnen geïnterpreteerd worden in het licht van een passend baseline corpus.

De wet van Zipf houdt onverminderd stand in de spraak van mensen met afasie, zij het met andere waardes van de parameters. Dit wijst erop dat zij woorden op dezelfde wijze in hun geheugen opslaan en hier weer uit halen als mensen zonder afasie. Het systeem *an sich* is dus ongewijzigd, alleen de bereikbaarheid van de opgeslagen woorden is beperkt. Dit maakt het aannemelijk dat de wet van Zipf zijn origine heeft in het mentale lexicon.

Curriculum Vitae

Marjolein van Egmond was born on December 22, 1986 in Gouda, and grew up in Rijswijk, Castricum and Heemskerk. After obtaining her VWO diploma, she studied *Dutch language and culture* at Utrecht University. She obtained her Bachelor's degree cum laude in 2009, with a major in linguistics. She continued with the research master *Linguistics: The study of the language faculty* at the same university. She majored in psycholinguistics and minored in phonetics, and obtained her Master's degree cum laude in 2011.

In 2011, she was awarded an NWO grant for an individual PhD project at the Utrecht Institute of Linguistics OTS. Between 2011 and 2018 she worked as a PhD researcher, of which this dissertation is the result.