MASTER'S THESIS

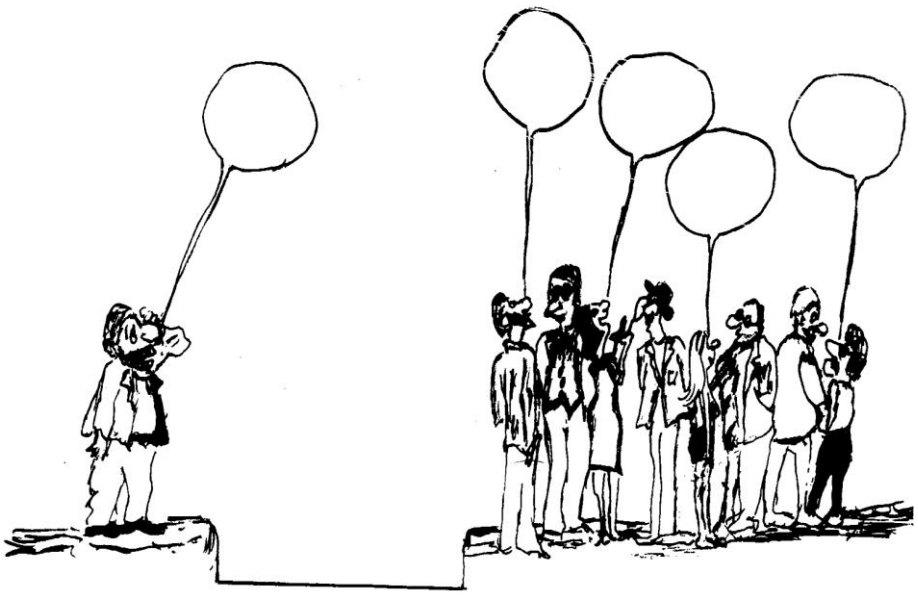
Word Finding Difficulties in Aphasia and their Effect on Zipf's Law

Marjolein van Egmond

Supervisor: Prof. Dr. Sergey Avrutin

Second Reader: Prof. Dr. Ted Sanders

July 1st, 2011



Skip Marcella - http://skipmarcella.shawwebspace.ca

Abstract

The current project combines two very different lines of research to investigate word finding difficulties of non-fluent aphasic patients in spontaneous speech. The first line is that of lexical retrieval in aphasic speech production. A literature review shows that the model by Levelt, Roelofs & Meyer (1999) can be used to accurately describe aphasic word finding difficulties. These difficulties likely arise due to reduced processing capacities. The amount of effort that is necessary to process words can be influenced by several variables, such as frequency, age of acquisition and entropy. The variable frequency provides the link to the second line of research, which is the distribution of word frequencies. The frequency distributions of natural language texts follow a power law called Zipf's law. Deviations from this law have been found for different groups of patients. In the current study, Zipf's law was investigated in the spontaneous speech of four non-fluent aphasic patients. Four speakers from the Corpus Gesproken Nederlands (Corpus Spoken Dutch) served as control group. Results show that speech from all patients conforms to Zipf's law. A difference between the two groups was found in the slope of this distribution, which is due to the fact that aphasic speech shows a less varied vocabulary and larger groups of high frequency words. This finding is explained as indicative of an unimpaired lexicon and an adaptation to reduced processing capacities. A detailed suggestion for future research is provided, in which the disruptions found in speech from aphasic patients is hypothesized to be reflected in the numbers of lexical connections of the words that occur in their spontaneous speech.

Table of Contents

1. Introduction

Language is one of the essential aspects of human communication. A disturbance of language has therefore a great influence on people's lives. Such a disturbance is the characteristic feature of aphasia, a language disorder acquired by almost 10,000 people in the Netherlands each year. Aphasia is caused by brain damage, such as a stroke, an accident or a brain tumor (Afasie Vereniging Nederland). Non-fluent aphasia, the main focus of the current study, is characterized by effortful, telegraphic speech, omission or substitution of functional elements and word finding difficulties.

In order to develop diagnostic tools and therapy it is necessary to have a solid understanding of the underlying cause of these problems. An influential hypothesis for the possible underlying cause is a reduction of the language processing capacities (e.g. Avrutin, 2006; Burkhardt, Avrutin, Piñango & Ruigendijk, 2008). The effect of these reduced capacities is that the information load that can be processed is lower: the brain has to adjust itself to be able to process information with limited resources.

The purpose of the current work is to investigate the effect of these word finding difficulties on Zipf's law, a law for word distributions in healthy speech. Word frequencies in a text or conversation show a specific distribution that is characteristic for human language. It will be investigated whether this distribution also applies to speech from aphasic patients.

First, in Section 2, a theoretical framework will be provided. Lexical retrieval in healthy speakers will be discussed, followed by a discussion of lexical retrieval in aphasic speakers and a discussion of the variables that influence lexical retrieval. This section is followed by a discussion of Zipf's law and its relevance to aphasia. The research questions for the current study are formulated in Section 3. The methods used to answer these questions are given in Section 4, followed by the results in Section 5. A discussion of these results, including detailed suggestions for future research, is given in Section 6.

2. Theoretical Framework

2.1 Lexical Retrieval in Healthy Adults

For a good understanding of the problems caused by aphasia it is necessary to have a solid understanding of lexical retrieval in healthy people. Here, solely content words will be examined. Function words are selected differently from content words: content words are selected on semantic grounds, while function words are selected on syntactic grounds (Levelt, Roelofs & Meyer, 1999). This difference is reflected in speech from especially non-fluent aphasic speakers: they are well-known for their omission or substitution of functional elements (e.g. Grodzinsky, 2000).

Several models of lexical selection have been proposed. One of the most influential models was developed by Levelt and colleagues. This model is based on the concept of spreading activation. According to Levelt, Roelofs & Meyer (1999), lexical access of content words is achieved through a serial two-system architecture. The first system is that of lexical selection, which consists of perspective taking and lemma selection. During perspective taking the speaker focuses on a lexical concept. The speaker has to estimate how much detail is wanted: if he wants to talk about a horse, he can use the word *horse*, but he could also use the word *stallion* or *animal*. During perspective taking there is co-activation of related concepts. Each active lexical concept spreads activation to the corresponding lexical item in the speaker's mental lexicon. The lemma that is eventually chosen is the one with the highest level of activation. This level of activation depends on the number of connections between the target word and other words, the strength of these connections and the initial level of activation. So the target lemma, the lemma that is eventually chosen, is selected under competition. The selection latency depends on the amount of co-activation of other lemmas.

The second system is that of form encoding, which consists of the retrieval of morphemic and phonological codes, prosodification and syllabification and phonetic encoding. According to Levelt, there is no competition in this phase: activation spreads from just the selected lemma to the phonological codes needed for its articulation, without other codes being activated.

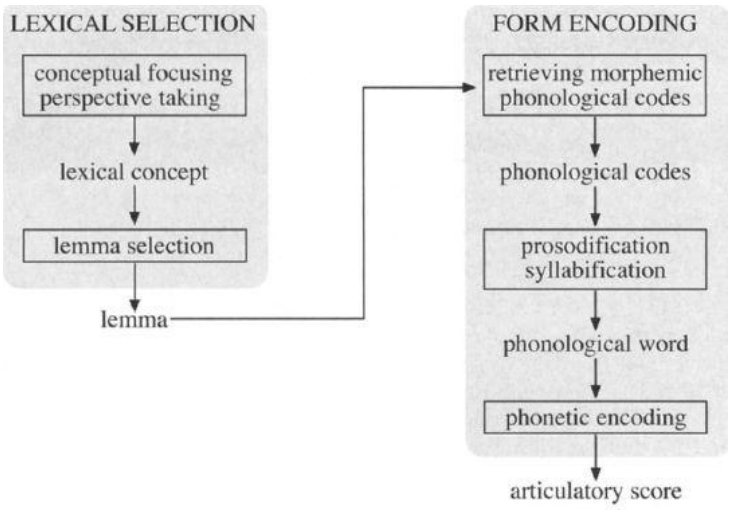


Figure 1. Levelt's model of lexical retrieval

Sahin, Pinker, Cash, Schomer and Halgren (2009) have tested whether the different phases of Levelt's model can be localized in the brain, both in time and space. They tested three epileptic patients, who, as part of a pre-surgical procedure, had deep electrodes inserted in Broca's area. Through these electrodes, local field potentials (LFP) could be recorded. Patients were given two tasks: they were asked to silently produce words or to silently fill in the blank, thereby inflecting a given target word. The individual items of both tasks were mixed, to prevent the patients from relying on task specific strategies. Half of the words required regular inflection while the other half required irregular inflection. Also, words in which the inflected form was phonologically identical to the stem (null-inflection) were included. Results indicated a clear three-step procedure of lexical retrieval, particularly in Broca's area (more specifically, Brodmann area 45). This three-step procedure could be assigned to slightly different locations inside Broca's area, but were more strongly identifiable in time. Word identification was performed about 200 ms after the stimulus was presented. An effect of word frequency was visible: stronger responses were measured for low frequency words compared to high frequency words. Next, morphological composition was performed around 320 ms after stimulus presentation, which was shown by a difference between tasks but not between words with overt or null-inflection. Phonological encoding took place around 450 ms after stimulus presentation: a difference was seen between words with overt inflection or their stem but not between words with null-inflection and their stem. The pattern was the same for both nouns and verbs. These findings provide strong evidence in favor of Levelt's serial model of lexical retrieval, not only as a purely theoretical model but also as an accurate description of physical lexical retrieval.

2.2 Lexical Retrieval in Aphasic Patients

Much is known about lexical retrieval in healthy adults. Much less is known about lexical retrieval in aphasic patients, even though aphasia was first described by Hippocrates (ca. 460 – ca. 370 B.C.) (Günter, Hofman & Promes, 2009).

Many aphasic patients suffer from word finding difficulties: they know what they want to talk about, they can provide a description or give related or associated words or draw what they mean, but they cannot retrieve the word itself. Word finding difficulties in aphasia can occur when lexical retrieval fails. The question remains, however, what causes this failure in the first place.

It has been widely accepted that many problems faced by aphasic patients are caused by processing difficulties. However, the nature of these difficulties is still subject of debate (Burkhardt, Piñango & Wong, 2003). According to Burkhardt et al. two approaches have been proposed: the *dependency-relations approach* and the *time-course approach*. These approaches are mainly concerned with the syntactic difficulties faced by many patients. The dependency-relations approach claims that syntactic difficulties arise during the implementation of certain kinds of dependency relations which arise after NP- or Wh-movement. This means that not only performance but also

competence is affected: aphasic patients suffer from a loss of knowledge. The time-course approach on the other hand claims that syntactic processing is slowed down. This slowing down can cause an overload of the processing system, which results in an inability to carry out syntactic processing at a normal rate. Processes that depend on a fully formed syntactic structure are hereby affected, such as the assignment of thematic roles. This means that performance is affected while competence is fully intact (for a detailed overview of both approaches the reader is referred to Caplan, Waters, DeDe, Michaud & Reddy, 2007).

One specific account of the time-course approach was put forward by Avrutin (2006). According to this theory, the effect of processing overload is that syntax is not always the most efficient way to encode information. If syntax is weakened due to brain damage then other systems – such as context – can be used to encode a message or to build information structure in comprehension. These alternative systems are in specific contexts also available to healthy speakers. An example is the Dutch sentence: "Marie vertelde Peter een mop: en hij *lachen*!" ("Mary told Peter a joke: he laugh-INF!") Only in a context like the first sentence is the second sentence allowed. Avrutin claims that aphasic patients also rely on this context-based system in cases where healthy speakers would rely on syntax.

Of these two approaches, only the second provides an explanation for word finding difficulties. If processing resources are reduced, then the brain has to adjust itself to be able to process information with limited resources. This can be done by reducing the complexity of lexical access. As explained above, during perspective taking several related lexical concepts are activated. The word that is eventually activated is the word with the highest level of activation. Lexical selection is more complex if other words have a level of activation that is close to the level of activation of the target word. A mathematical measure of this complexity is entropy. This measure can be used to calculate the degree of complexity of lexical retrieval (De Lange, 2008). Higher entropy requires higher processing capacity, which in healthy participants results in longer response latencies (Baayen, Levelt, Schreuder & Ernestus, 2007).

In principle, the inability to produce a target word could be due to a breakdown at any phase of word retrieval. Whether Levelt's two-stage model is able to account for the problems encountered by aphasic patients has been investigated by Laine, Tikkala & Juhola (1998). They used a computer model of Levelt's two-stage architecture and tested whether it was able to account for the specific problems of ten Finnish aphasic patients. This group of aphasics was very heterogeneous: Two of these patients were diagnosed as Broca's aphasics, three as Wernicke's aphasics, two as conduction aphasics and three as anomic aphasics. Their model consisted of two separate but connected networks that corresponded to the two systems of Levelt's model, namely a lexical-semantic network and a phoneme network. Errors were produced by manipulating noise ratios in both networks and the selection thresholds for the lexical-semantic network. The results showed that the model was successful in simulating the error types and error rates of the different patients. These results locate

word finding difficulties at the first system of Levelt's model (noise ratios), or at the ridge between both systems (threshold values).

2.3 Processing Cost

If the time-course approach discussed above is correct then the words that are difficult for aphasics to retrieve are those words that require too much effort to process. But which factors determine the required amount of effort for a word?

This question was mainly addressed from two angles: by focusing on speech errors and by focusing on response latencies. Kittredge, Dell, Verkuilen & Schwartz (2008) took the first approach. They performed two separate regression analyses to explore the relationship between targets and errors in aphasic patients. Their analyses were based on picture naming data from 50 aphasic patients suffering from a diverse range of impairments. Their first analysis was aimed at determining which lexical properties of the target make it more or less susceptible to errors, while the second was aimed at exploring the relationship between the target's lexical properties and those of the error it elicits. They found that high target log[1] frequency predicted fewer semantic, phonological and omission errors and a higher log frequency of the error words. The effect was strongest for phonological errors. These findings suggest a frequency effect on both lexical selection and form encoding.

These findings contradict the traditional view that frequency effects solely exist at the lexeme level, which is the first level of form encoding. This view was most prominently set forward by Jescheniak & Levelt (1994) and is called the Word Frequency Effect. Their most important finding was that low frequency words with a highly frequent homophone were translated equally fast from English to Dutch as high frequency words. Homophones share their lexeme (abstract phonological form) but not their lemma (semantic properties and sub-categorization features) or meaning. Therefore, an effect of a highly frequent homophone on the translation of a low frequent word suggests a frequency effect at the lexeme level.

Kittredge et al. (2008) were not the first to challenge the traditional view on the Word Frequency Effect. Their review of previous literature shows that although some studied confirmed a frequency effect on the phonological level (e.g. Dell, 1990; Laubstein, 1999; Vitevitch, 1997), other studies have shown a frequency effect for lexical retrieval as well (e.g. Gahl, 2006; Alario, Costa & Caramazza, 2002, Harley & MacAndrew, 2001). The studies reviewed by Kittredge et al. (2008) and the studies discussed above suggest that frequency influences lexical retrieval at several stages, although the effect might be most pronounced at the lexeme level.

[1] Log-transformed frequency values were used, because log frequency was previously found to be better correlated with many measures of performance.

Meanwhile, the existence of a frequency effect, irrespective of the level at which it originates, remains widely accepted. The effect of frequency for aphasic patients was shown by Cuetos, Aguado, Izura & Ellis (2002). They had 16 Spanish-speaking aphasic patients perform a naming task and measured naming accuracy. Their results show that words are easiest to retrieve if they are acquired at a young age, if they are familiar to the patient and if they are highly frequent. Visual complexity, imageability, animacy and word length were not significant[2]. This shows that frequency, but also age of acquisition and familiarity influence lexical retrieval by aphasic patients.

Recently, Baayen, Levelt, Schreuder & Ernestus (2007) performed an investigation of the factors influencing ease of lexical retrieval using response latencies. They did two picture naming experiments in which line drawings of singulars and plurals were presented to healthy participants. The first experiment required direct responses; the second experiment required delayed responses. Half of the target nouns were high frequent words while the other half were low frequent words. For each frequency level half of the nouns was singular dominant while the other half was plural dominant (a variable called dominance). All singulars were monosyllabic; all plurals were bisyllabic. The first experiment showed that low frequency words elicited longer response latencies than high frequency words. An interaction effect with dominance was found: plural dominance gave rise to a processing disadvantage that extended to both the singular and the plural form. For all words entropy was calculated. This measure provided an even better predictor of response latencies than dominance, with higher entropy resulting in longer response latencies. Relative entropy, a measure for the extent to which the probability distribution of a particular noun diverges from the corresponding probability distribution of the class of nouns, also had an inhibitory effect. The effect of number of meanings as measured in number of Synsets in the WordNet database (Miller, 1990; Fellbaum, 1998) was most prominent for higher Synset counts, for which it was inhibitory: more meanings resulted in longer response latencies. These factors did not reach significance in the delayed picture naming task, which means that the origin of their effect lies in lexical retrieval rather than in articulation. In sum, lexical retrieval of words with higher lexical complexity as measured by entropy and number of Synsets takes longer, which implies higher processing effort. Their origin can be located in the first system of Levelt, Roelofs & Meyer's model.

Baayen et al. (2007) attempted to replicate their first experiment using photographs in stead of line drawings. In this experiment they tested singular, dual and plural nouns. For this experiment results were less clear: a facilitatory effect of relative entropy was found for plurals but not for singulars and duals, and an inhibitory effect of entropy was found for singulars and duals but not for plurals. Baayen et al. suggest that this might be due to the difference between numerical specificity.

[2] Visual complexity, imageability, animacy and word length did predict naming accuracy in at least two individual patients, but for the group as a whole they were not significant.

A comparison between the effect of information theoretical measures and more traditional measures on lexical retrieval was performed by Moscoso del Prado Martín, Kostiç & Baayen (2004). They formulated an information residual measure, which was calculated as the difference between the amount of information contained by a word (more information causes higher processing demands) and its total paradigmatic entropy (an estimate of the facilitatory effect of the morphological paradigms to which a word belongs). This measure was then compared to surface and base frequency, family size and cumulative root frequency in its ability to predict response latencies in three visual lexical decision tasks performed in previous experiments. Results show a facilitatory effect of frequency and of morphological family size and in two experiments (when the effect of morphological family size is partialled out) an inhibitory effect of cumulative root frequency. Also, their information residual measure outperforms the other measures in terms of explained variance.

Results from comprehension experiments do not necessarily extend to production experiments. In fact, variables that are facilitatory in production experiments can be inhibitory in comprehension experiments (Baayen, 2007). What we can conclude, however, is that information theoretical measures can play an important role in explaining response latencies in lexical retrieval.

The role of information theoretic measures in aphasic speech production has not yet been tested. In comprehension, one study is currently being carried out by Van Ewijk & Avrutin (personal communication). In this project van Ewijk & Avrutin aim to explore the complexity of individual verb forms and its effect on lexical retrieval, both in healthy and aphasic people. Preliminary results show that inflectional entropy (reflecting complexity within a given verbal paradigm) is a significant predictor for response latencies for both healthy elderly adults and aphasic patients. However, linear mixed model analyses showed an interaction between group and inflectional entropy, indicating that for aphasics the effect of inflectional entropy was less pronounced. Their findings provide the first encouraging results that the characteristics of lexical processing capacity for aphasic patients can be calculated using information theoretical means.

Steyvers & Tenenbaum (2005) investigated the effect of frequency, age of acquisition and semantic density from a very different angle: they performed a statistical analysis of the large-scale structure of 3 types of semantic networks: word associations, WordNet and Roget's Thesaurus. This analysis shows that the distribution of the number of connections follows power laws: most words have relatively few connections, and they are joined together through a small number of words with many connections. They suggest that his organization is the result of the way in which semantic networks grow. Steyvers & Tenenbaum present a model of semantic growth to support this claim. In this model each new word is connected to the existing network by differentiating the connectivity pattern of an existing node. This method resulted in more frequent, early acquired words showing higher connectivity.

In sum, the results of the studies above show the following. Word frequency, familiarity and age of acquisition have a facilitatory effect on lexical retrieval in production. High lexical connectivity as measured by entropy or number of Synsets has an inhibitory effect on lexical retrieval in production.

As discussed in section 2.1, the level of activation that is eventually reached depends on the number of connections a word has, on the strength of those connections and on their initial level of activation. Therefore, an investigation of aphasic word finding difficulties should start with an investigation of these three variables. An important question is how the variables discussed above influence these three factors.

A possible explanation runs as follows. The factors likely to be influenced by frequency are connection strength and the initial level of activation. Mental connections that are frequently used are likely to become well established in the brain. Also, a high initial level of activation allows frequent words to be easily retrieved. Age of acquisition is likely to influence the number of connections a word has: words that are newly learned connect to previously learned words, causing the early acquired words to function as highly connected hubs in the lexical network. Strong connections, high initial levels of activation and high connectivity allow these words to be easily retrieved, which gives a feeling of high familiarity. This results in a facilitatory effect of word frequency, familiarity and age of acquisition. But not only the target word plays a role in lexical retrieval. As discussed in section 2.1, during perspective taking activation spreads from the mental concept to the target word but also to words related to the target. The word that is eventually activated is the word with the highest level of activation. This explains the inhibitory effect of number of Synsets: a higher number of Synsets means that more words become co-activated, which renders it increasingly difficult to select the target word. Lexical selection is more complex if other words have a level of activation that is close to the level of activation of the target word. This complexity is measured in entropy, which explains the inhibitory effect of entropy measures.

2.4 Zipf's Law

Above, it was shown that word frequency has a facilitatory effect on lexical retrieval. It seems plausible to assume that those words that are frequent in a language in general are also more frequent in speech from a single person. Words that are frequently encountered in a language are frequently accessed in the lexicon. This renders highly frequent words to be easy to access and therefore to be used frequently.

In the early sixties of the previous century, George Kingsley Zipf showed that word frequencies in natural texts follow power laws: the frequency of any word is inversely proportional to its rank (Zipf, 1965). In other words, the most frequent word (rank 1) in a text will occur approximately twice as often in a text as the second most frequent word (rank 2), three times more often than the third most frequent word (rank 3), etc. This distribution results in a linear dependency between rank and frequency if the data is plotted on a doubly logarithmic scale. An example of such a distribution on both a normal and a doubly logarithmic scale is shown in Figure 2.

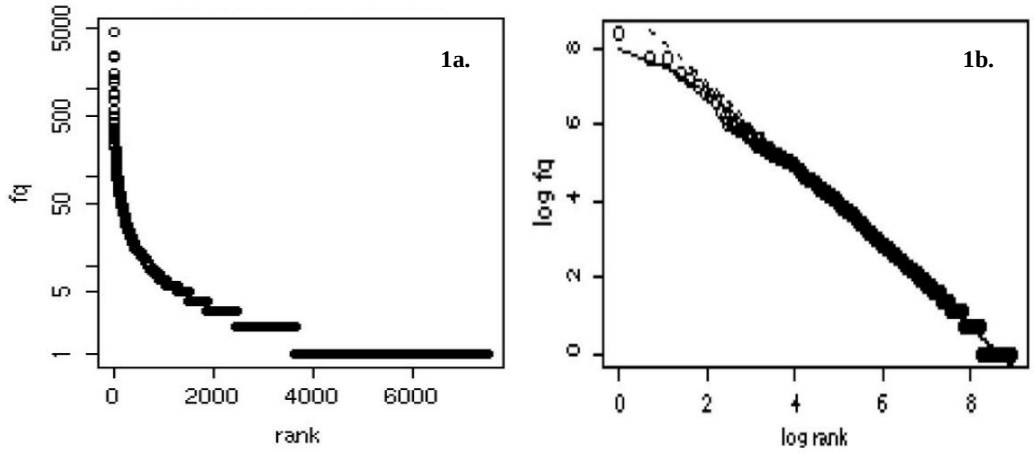


Figure 2. Zipf's law in H.G. Well, *The War of the Worlds* (1989) (Baroni 2008). Figure 1a shows a rank/frequency plot on a normal scale, figure 1b shows the same rank/frequency plot on a log-log scale.

Zipf's law has been shown to occur in many other ratings, some of which are unrelated to language. Examples are populations in city sizes, firm sizes in industrial countries or family names (e.g. Corominas-Murtra & Solé, 2010; Dahui, Menghui & Zengru, 2005). According to Zipf (1965, p.1) this distribution follows from a Principle of Least Effort:

> "[T]he structure and organization of an individual's entire being will tend always to be such that his entire behavior will be governed by [the] Principle [of Least Effort]."

He defines this principle as follows:

> "In simple terms, the Principle of Least Effort means, for example, that a person in solving his immediate problems will view these against the background of his probable future problems, as estimated by himself. Moreover he will strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems. That in turn means that the person will strive to minimize the

> probable average rate of his work-expenditure (over time). And in so doing he will be minimizing his effort, by our definition of effort." (p. 1)

Corominas-Murtra & Solé (2010) account for the emergence of Zipf's law in a more precise manner. They show that Zipf´s law is the inevitable outcome of a general class of stochastic systems:

> "[… W]e treat complex systems as stochastic systems describable in terms of algorithmic complexity and thus statistical entropy. […W]e define a characterization of a wide class of complex systems, which grasps the open nature of many complex systems, summarized in [a single equation]. The main achievement of this equation is that it encodes the concepts of growing and, even most important, the stabilization of complexity properties in an intermediate point between order and disorder, a feature observed in many systems displaying Zipf's-like statistics. From this equation we derived Zipf's law as the natural outcome of systems belonging to this class of stochastic systems." (p. 6)

How this process affects language is examined by Ferrer i Cancho & Solé (2003). They show using a mathematical model[3] that Zipf's law is the outcome of the nontrivial arrangement of associations in a lexicon that has to comply with hearer and speaker needs[4]. The speaker would prefer to use as few different words or signals as possible to express a message (unification), whilst the listener would prefer a different word for every meaning (diversification). Theoretically, speaker effort is minimal if one word is used to express all meanings: in this case the effort of lexical search would be reduced to the retrieval of a single, easily accessible word. But this would maximize hearer effort due to maximal ambiguity. Theoretically, hearer effort is minimal if every meaning is expressed with a different word, because this would minimize ambiguity and therefore uncertainty. Zipf's law would then occur as a result of the tension between these two needs: the vertical part in a graph like Figure 2a (until about 1000 words) results from the speaker need of unification, while the horizontal part of the graph (from about 2000 words onwards) results from the hearer need of diversification.

The finding that Zipf's law applies to all natural language texts is not trivial. As Popescu, Altmann & Köhler (1999) formulate it:

> "[I]f a super-system such as a language organizes itself according to some principle such as word cost optimization then why should exactly the probability distribution which results from an overall optimization be found in every individual text, where deviations from the structure of the super-system would not do any harm? We should be able to find at least some texts in which rare words occur rather often and frequent words are infrequent or even absent and which display a form considerably deviates from the Zipf-like shape.
>
> In reality however, all texts in all languages seem to conform to the Zipfian shape more or less closely (…)."

[3] More details of this model are provided in Section 6.1.

[4] This includes the needs of different conversational partners but also the needs of a single person who in communication alternates between hearing and listening.

It has been argued that Zipf's law does not reflect underlying properties of stochastic systems, but rather result from random processes. However, Ferrer i Cancho and Elvevåg (2010) have shown that frequency distributions of random texts do not resemble the frequency distributions of natural language. For this, they compared ten random texts generated by different processes and with different parameter settings to ten English texts. In all cases, the random texts turned out to be statistically inconsistent.

Speech from Aphasic patients is markedly different from normal speech. Especially speech from non-fluent patients is telegraphic and effort-full. Whether these problems are also reflected in the frequency distributions of their speech has not yet been tested.

Piotrovskii, Pashkovskii & Piotrovskii (1994) and Piotrowski and Spivak (2007) did study 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; for schizophrenic patients with a topic of obsession they found a curve in stead of a straight line, and for children with Down syndrome they found a steeper slope. These findings suggest that differences might be found for aphasic speakers as well.

Usually, Zipf's law is studied for all words of a text. The focus for the current study, however, is on content words only. That Zipf's law also applies to different components of natural language texts has been shown by Popescu, Altmann & Köhler (2010). They fitted two mixed exponential components to the frequency distributions of 100 texts in 20 languages. In only three of the 100 texts a better fit was provided by the single component of Zipf's law. Popescu et al. show that this finding is due to the difference between the distribution for synsemantic words and autosemantic words, roughly the distinction between function words and content words. This study shows that Zipf's law can be studied for content words only, and that in fact a better fit might be found to content words only then to the text as a whole.

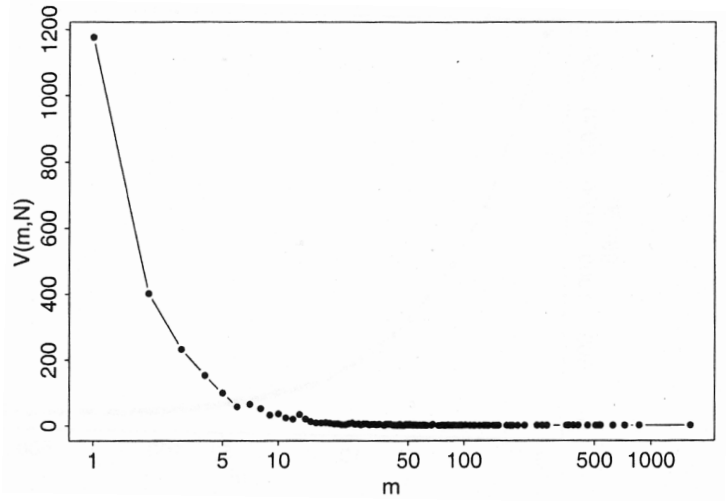


Figure 3. The frequency spectrum of Alice in Wonderland (m: frequency class; V(m,N): number of types with frequency m) (Baayen, 2001).

Estimation of the parameters of Zipf's law is not straightforward. Linear regression in doubly logarithmic scales can lead to a highly biased estimation of the exponent if data points do not follow a normal distribution (Ferrer i Cancho, personal communication). One way to circumvent this problem is by studying the frequency spectrum instead of the frequency distribution, of which an example is given in Figure 3. The frequency spectrum shows the number of types *V(m,N)* in each frequency class, where a frequency class is the number of words occurring exactly *m* times in a sample of *N* words in total (Baayen, 2001: 10). This means that the group of data points for each frequency is reduced to one data point per frequency class. The result of this is that each frequency class is assigned equal weight, which is not the case for the frequency distribution where lower frequency classes receive more weight due to the larger number of data points in these classes. A disadvantage of this method is that it involves quite severe data reduction. In what follows, Zipf's law will therefore be examined through frequency spectra in stead of frequency distributions.

3. Research Questions

The literature review in Section 2 has shown that word frequency has a facilitatory effect on lexical retrieval. This effect has also been shown for aphasic speakers. Possibly, this effect is due to the higher lexical connection strength of words that are frequently used. A different field of research as shown that word frequency distributions conform to a power function called Zipf's law. Deviations from this law have been found for different groups of patients. In line with these findings two research questions were formulated:

1. **Does the frequency spectrum of spontaneous speech of aphasic speakers follow Zipf's law?**
2. **If the frequency spectrum of aphasic speakers conforms to Zipf's law, then is the slope of Zipf's law different for aphasic speakers compared to healthy speakers?**

From Section 2.2 it followed that lexical retrieval in aphasia fails if processing costs become too high. The studies discussed in Section 2.3 showed that word frequency had a facilitatory effect on lexical retrieval. It is therefore expected that aphasic speakers use more high frequency words and less low frequency words. This would result in a shallower slope of Zipf's law in the frequency spectra of aphasic speakers compared to healthy speakers.

4. Methods

4.1 Participants

Four aphasic speakers were recruited. The patients JvdH and PH were recruited from Afasiecentrum Tilburg, a day care centre for aphasic people; EvdL and JJ were recruited from Samen Verder in Tilburg (although both also attended day sessions at Afasiecentrum Tilburg), a private evening group of aphasic patients who want extra practice and social contact with other aphasic patients. All patients were diagnosed as non-fluent by speech therapists. All patients were at least 3 years post onset. Details of the aphasic speakers are given in Table 1.

As a control group, four healthy speakers were selected from the Corpus Gesproken Nederlands (CGN, Nederlandse Taalunie, 2004). They were matched on sex with the aphasic speakers. Two recordings were chosen: fn000260 and fn000276. Both recordings contained spontaneous conversations of two people of sufficient length, each of which provided a match with one of the aphasic speakers. Recording fn000260 contained speech from the speakers N01004 and N01005; recording fn000276 contained speech from the speakers N01010 and N01011.

Speaker details are given in Table 2.

Table 1. Details Aphasic Speakers

	Time post onset	**Cause**	**Type of aphasia**
EvdL	10 year, 4 months	Stroke	non-fluent (Broca)
JJ	7 year, 3 months	Multi-infarct syndrome	non-fluent (Broca)
JvdH	6 year, 3 months	Stroke	non-fluent (Broca)
PH	3 year, 10 months	Stroke after trauma	non fluent (Broca)

Table 2. Speaker Details. The CGN only provides an age range, not the exact age at the time of recording.

Healthy speakers			**Aphasic speakers**		
Speaker	**Sex**	**Age**	**Speaker**	**Sex**	**Age**
N01011	female	25-34	EvdL	female	59
N01005	female	56 or older	JJ	female	63
N01004	female	25-34	JvdH	female	36
N01010	male	25-34	PH	male	33

4.2 Procedure

4.2.1 Interviews

Spontaneous speech from aphasic participants was obtained through unstructured interviews which lasted for about 20 minutes per participant. These interviews took the form of informal conversations, which means that interviewers actively participated in the conversation. This approach was chosen for two reasons. First, active participation is often the only way to establish mutual understanding. On the one hand, by reformulating the patient's utterances the interviewer is able to check whether she correctly understood the patient. On the other hand, patients are often insecure whether they are understood correctly. Reformulation by the interviewer signals the interviewer's degree of understanding and allows for corrections. Second, the goal for the current project was to elicit spontaneous speech. An informal conversation provides a natural situation for spontaneous speech, while a strict question-answer format would create an unnatural situation. Such an unnatural situation is likely to influence the spontaneous speech: patients would feel more uncomfortable, causing them to perform worse than normal.

4.2.2 Analysis

All interviews were recorded on video. They were orthographically transcribed using the CHAT-format and labelled for part of speech. The CHAT-format is part of the CHILDES project (MacWhinney, 2000). This format allows for automatic searches and analyses by means of the accompanying CLAN program.

The speech fragments that were selected from the CGN were manually converted from PRAAT-format to CHAT-format.

For all speakers, the first 102 content words (nouns, verbs, adjectives and adverbs) were selected for analysis. This number is equal to the number of content words in the smallest sample (EvdL).

Details about the statistical analysis that was performed will be provided in the results: this way it will be clearer why certain choices were made.

5. Results

The frequency distribution of healthy speakers and aphasic speakers on a doubly logarithmic scale is given in Figure 4. Visual inspection of this distribution shows that the data closely follows a straight line, indicating that Zipf's law applies.

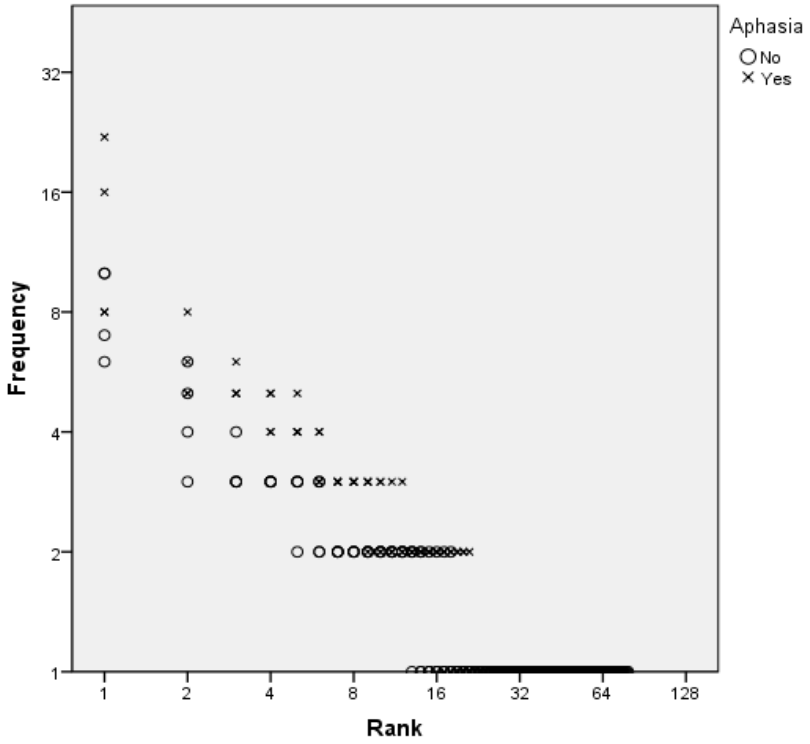


Figure 4. Frequency Distribution

Next, the frequency distribution was converted into a frequency spectrum. This involved a severe data reduction: The participant data was reduced from 102 data points to on average six data points per participant. The group data was reduced from 408 to 20 (healthy speakers) or from 408 to 25 (aphasic speakers) data points. This reduction causes only highly significant effects to surface.

Both frequency classes and *V(m,N)* were transformed to their logarithms. Only frequency classes in which words were present were included in the data, which excludes the possibility of empty frequency classes.

The frequency spectrum, shown in Figure 5, shows a straight line with a negative slope when plotted on a doubly logarithmic scale. For both groups, frequency class is a significant predictor of the size of the frequency class. The model provides a good fit for the data: R^2 for aphasic speakers is 0,888; R^2 for healthy speakers is 0,859. Notice that the data points do not need to follow a negative sloping line: for example, different frequency classes could have had equal sizes, which would have resulted in a non-sloping line. This negative sloping line shows that Zipf's law applies.

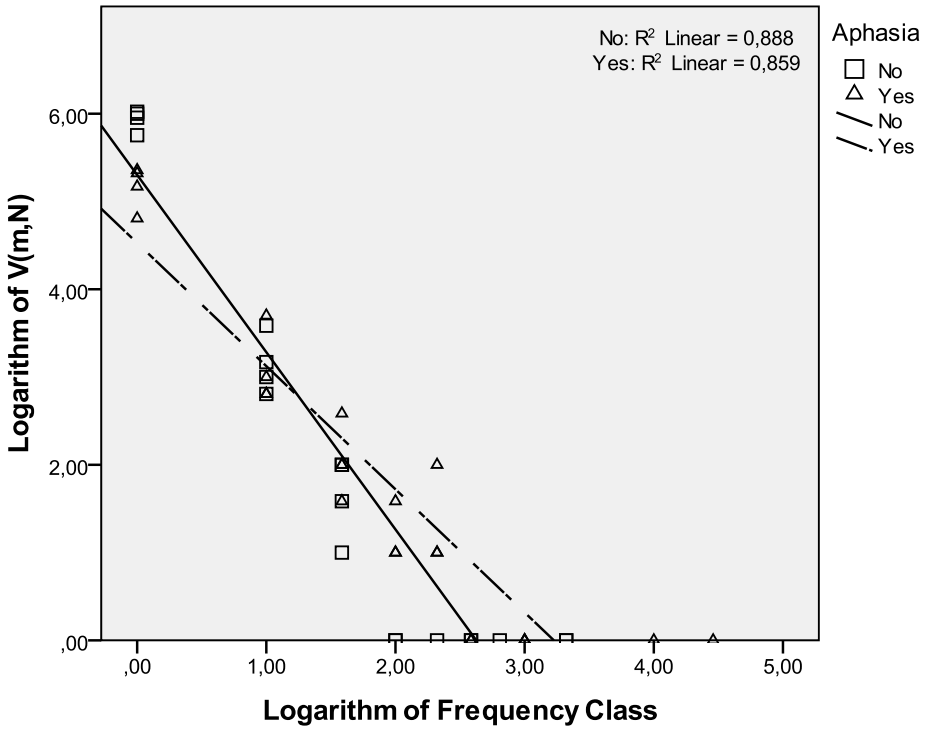


Figure 5. Frequency Spectrum

The frequency spectra for the individual speakers are given in Figure 6. For each speaker, frequency class is a significant predictor of the size of the frequency class. For all speakers, $R^2 \geq 0{,}830$ which indicates a good fit of the model. Even though these findings are expected if Zipf's law applies, it is no obvious finding: sample sizes were very small after the transformation from frequency distributions to frequency spectra. Details about slope values and their statistical significance are given in Table 3. A visual representation of slope values with their standard error ranges is given in Figure 7. In all cases the slope is negative: for every speaker low frequency classes were largest while high frequency classes were smallest.

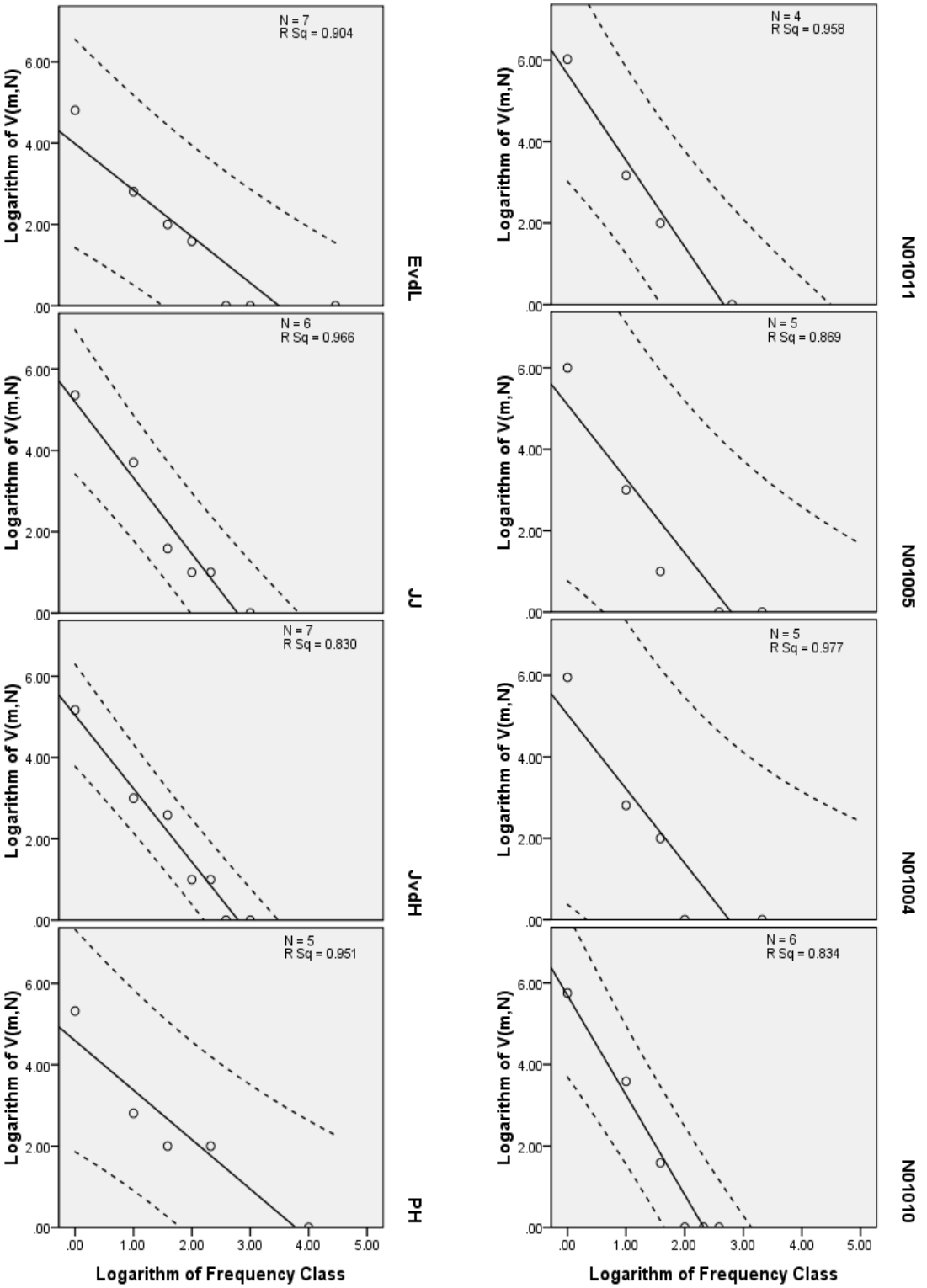


Figure 6. Frequency Spectrum per participant after axis-transformation

Table 3. Slope values and statistical details

Participant	N	Per Participant	Per Group
EvdL	7	**-1,140** *(95% CI: -1,733 to -0,547)* t(6) = -4,942; p < 0,01; r^2 = 0,830	**-1,403** (*95% CI: -1,648 to -1,158*) t(24) = -11,859; p < 0,01; r^2 = 0,927
JJ	8	**-1,865** (*95% CI: -2,450 to -1,279*) t(7) = -8,846; p < 0,01; r^2 = 0,951	
JvdH	7	**-1,802** (*95% CI: -2,193 to -1,412*) t(6) = -11,874; p < 0,05; r^2 = 0,966	
PH	5	**-1,216** (*95% CI: -1,945 to -0,486*) t(4) = -5,301; p < 0,05; r^2 = 0,904	
N01011	4	**-2,121** (*95% CI: -3,119 to -1,124*) t(3) = -9,152; p < 0,05; r^2 = 0,977	**-2,019** (*95% CI: -2,374 to -1,664*) t(19) = -11,950; p < 0,01; r^2 = 0,942
N01005	5	**-1,821** (*95% CI: -3,123 to -0,520*) t(4) = -4,453; p < 0,05; r^2 = 0,869	
N01004	5	**-2,827** (*95% CI: -3,322 to -0,332*) t(4) = -3,889; p < 0,05; r^2 = 0,834	
N01010	6	**-2,448** (*95% CI: -3,162 to -1,734*) t(5) = -9,514; p < 0,01; r^2 = 0,958	

The 0,616 difference in slope between aphasic speakers and healthy speakers turns out to be highly significant (SE = 0,201; 95% CI = 0,212 to 1,020; two-tailed t-test; t(43) = 3,072; p < 0,01). This means that frequency class is a significant predictor for frequency class size in both groups. However, the rate with which class size changes is higher in healthy speakers than in aphasic speakers. When the number of tokens is kept constant, speech from healthy speakers contains less frequency classes than speech from aphasic speakers. Low frequency classes in aphasic speech are smaller than low frequency classes in healthy speakers. High frequency classes are larger in aphasic speech than in healthy speech. Also, speech from aphasic patients contains more high frequency classes than speech from healthy speakers.

These frequency spectra are a result of the less varied vocabulary used by aphasic speakers when compared to healthy speakers: aphasic speakers use less different words, which is reflected in the fact that they use less words only once or twice. Low frequency classes are thus smaller.

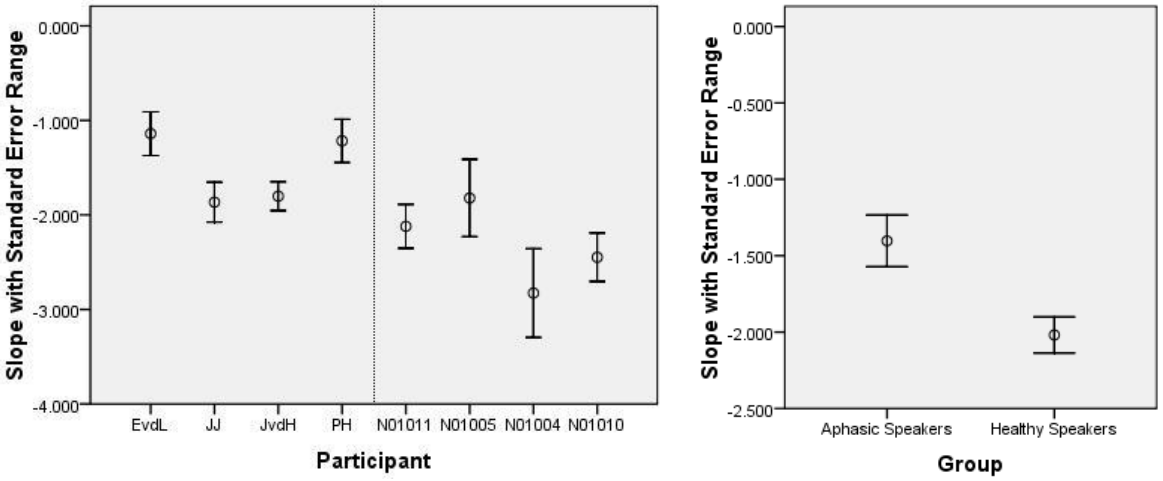


Figure 7. Slope with Standard Error Range per Participant and per Group (notice the different scale on the y-axes)

6. Discussion

The results show that the frequency spectra from the four non-fluent aphasic speakers that were investigated all follow Zipf's law. This answers the first research question. The second question was whether they would display a different slope on the frequency spectrum. The current results show that this is the case: aphasic speakers display a significantly more gradual slope than healthy speakers. This finding reflects the less varied choice of words of aphasic speakers. Aphasic speakers use fewer word types, but use those types more frequently.

This finding falls in line with the explanation of why Zipf's law occurs in the first place. As was discussed in Section 2.4, speaker effort is reduced if the number of different words used is as low as possible (Ferrer i Cancho & Solé, 2003). In Section 2.3 it was argued that aphasic word finding difficulties are caused by reduced processing capacities (e.g. Avrutin, 2006). The brain has to adapt to be able to communicate with these reduced capacities. For this, speaker effort has to be reduced. A likely way to achieve this is by a process of unification of word meanings: for some words lexical retrieval fails, so those words that can be retrieved from the lexicon are used to express a broader range of meanings. The effect of this on a frequency distribution would be more and larger high frequency classes and smaller and few low frequency classes. This is exactly what was found for the current data.

The finding that the speech from the aphasic speakers follows Zipf's law is remarkable, because their speech sounds clearly impaired. A short fragment from one of the interviews is shown on the next page to illustrate this (rough translations are provided in italics). The current results show that this impairment does not result in a different distribution of content words, except for a difference in slope. The abstract relationship between different frequency classes remain the same. In Section 2.4 it was argued that Zipf's law follows as output of a complex system, in this case the lexicon. The current finding that aphasic speech conforms to Zipf's law therefore suggests that the basic organisation of the lexicon is the same for both groups. This finding is in line with theories assigning aphasic word finding difficulties to lower processing rates rather than defects of the system. Lower processing rates would result in a relative preference for words that are easier to access but would not result in severe disruptions of frequency relations. The current results show exactly this.

The current results were obtained from only four aphasic speakers, who were compared to only four healthy speakers. Therefore, every individual speaker has a large influence on the results. Future research should try to replicate these findings to see whether they reflect true characteristics of non-fluent aphasic speakers or characteristics of the participants in the current study.

	(…)
Interviewer:	hee, wat doe je verder nog meer, behalve jeux-de-boules?
	hee, what else do you do, besides boules?
PH:	jeux-de-boules, uhm zw- zwemmen.
	boules, uhm, sw- swimming.
Interviewer:	oke [knikt ja].
	OK [nodds yes].
PH:	e:n… ja… uhm ma- uhm ma- maken, ook leu- ma-, hoe-heet-dat...
	a:nd… yes… uhm ma- uhm ma- making, also fu- ma-, how-is-it-called…
Interviewer:	oke, nog een keer, die snapte ik niet.
	OK, again, I didn't get that.
Interviewer:	iets maken.
	making something.
PH:	ja maken, gewoon.
	yes, just making.
PH:	uhm hout.
	uhm, wood.
Interviewer:	aah.
	aah.
PH:	of uhm ma- maken, gewoon, of wat...
	or uhm, ma- making, just, or what…
Interviewer:	beetje dingen maken met je handen.
	bit of making things with your hands.
PH:	ja kijk kijk maar [toont handen].
	yes, look, just look [shows his hands].
Interviewer:	oh ja [lacht], vandaar.
	oh yes [laughs], I see.
PH:	ja xx dus ja [lacht].
	yes xx so yes [laughs].
Interviewer:	beetje hout bewerking, dat soort dingen..?
	a bit of woodworking, that sort of things?
PH:	uhm ja.
	(…)

Even though the speech from the aphasic speakers sounded severely impaired, only gradual differences were found for their frequency spectra: the slopes differed, but the general distribution of tokens over frequency classes remained the same. The question remains, then, which variables would be affected by aphasic speech impairments. In what follows, one hypothesis will be put forward, which concerns the numbers of connections words have, measured by their age of acquisition. It remains a task for future research to test whether this hypothesis holds.

6.1 The Effect of Age of Acquisition on Lexical Retrieval

In Section 2.3 it was argued that the number of lexical connections a word has is likely to be determined by the age at which a word was acquired. Words that are newly learned connect to previously learned words, causing these early acquired words to function as highly connected hubs in the lexical network. This hypothesis was supported by the model of lexical growth developed by Steyvers & Tenenbaum (2005).

A model for the organization of numbers of lexical connections was developed by Ferrer i Cancho & Solé (2003). In this model, the lexicon is represented by a binary matrix. Each column in this matrix represents a word. All words together are called set S. This set of words is said to consist of *n* words. These columns represent the phonological form or PF of a language. Each row in this matrix represents a basic ingredient for word meaning, which Ferrer i Cancho and Solé call objects of reference or simply objects. All objects together are called set R. This set of meanings is said to consist of *m* objects. Together, they represent the logical form or LF of a language. A cell in this matrix contains a 1 if the word represented by its column has a lexical connection with the object represented by its row. Otherwise it contains a 0, which means that the word represented by its column has no lexical connection with the object represented by its row. Synonymy is allowed: one column can have 1's for several unrelated objects, which means that one word can be connected to several unrelated meanings. An example of a matrix is given in Table 4.

Table 4. Example of Matrix that Represents the Lexicon

		Phonological Form								
		Word 1	**Word 2**	**Word 3**	**Word 4**	**Word 5**	**Word 6**	**Word 7**	**Word …**	**Word n**
Logical Form	**Obj. a**	1	1	0	0	0	0	1	0	0
	Obj. b	1	0	0	1	0	0	0	0	1
	Obj. c	1	0	1	0	0	0	0	1	0
	Obj. d	0	1	0	1	0	0	0	0	0
	Obj. e	1	0	1	0	0	0	0	0	0
	Obj. f	1	1	0	0	1	0	0	0	0
	Obj. g	1	0	0	0	1	0	0	0	0
	Obj. …	0	0	1	0	0	0	0	0	0
	Obj. m	1	1	0	0	0	1	0	0	0

Ferrer i Cancho & Solé applied an algorithm to this model to distribute the 1's over the cells of matrix. This algorithm stopped if both speaker and hearer entropy were as low as possible. Speaker entropy was as low as possible if all 1's were assigned to only a few words: this would imply that only

a few highly ambiguous words would have to be retrieved from the lexicon to communicate every thinkable message. But this high ambiguity would maximize hearer entropy. Hearer entropy was as low as possible if the 1's were evenly distributed over all words: this would mean that distinct meanings are expressed by different words, thereby minimizing ambiguity. The outcome of this algorithm was that speaker and hearer entropy were lowest if the distribution of 1's over the matrix – and thereby the distribution of the number connections – followed Zipf's law.

These findings suggest that not only word frequencies but also number of connections follows a Zipfian distribution. But while word frequencies can simply be counted; the number of connections words have has to be derived. As discussed above, age of acquisition would provide a likely candidate for this derivation. From the model by Steyvers & Tenenbaum (2005) it follows that the relation between age of acquisition and number of connections is likely to be logarithmic. Such a relation would cause some early acquired words to become highly connected while most words have only a few connections, which is what they found for their model of semantic growth. If this is true then the distribution of number of connections can be derived through age of acquisition. For healthy speakers, this distribution is expected to conform to Zipf's law.

The question is, now, what the distribution would be for aphasic speakers. For this, the theory of lexical retrieval needs to be developed a bit further.

A target word is selected for lexical retrieval if the level of lexical activation of this word reaches or exceeds a critical threshold value. As was explained in Section 2.2, the energy for reaching this threshold spreads from the active mental concept. So the difference between this threshold value *T* and the level of activation of a word after the additional activation spreading from the mental concept has to nonexistent or positive:

[1] $$T - (y + \frac{A}{x}) \leq 0$$

Here, A is the level of activation stemming from the mental concept. Activation spreads not only to the target word, but also to related lexical entries or neighbours. This number of lexical neighbours is called *x*. The base level of activation is expressed by *y*. The difference between the threshold value and the sum of the base level of activation and the additional activation coming from

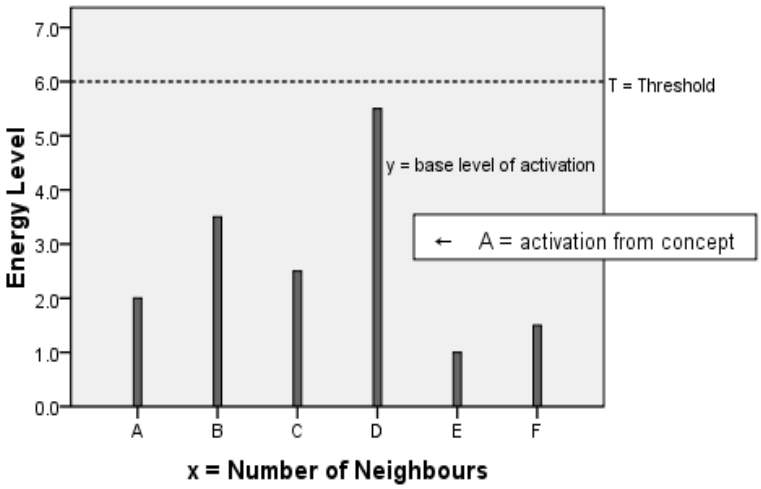


Figure 8. Graphical representation of the level of activation of a lexical family before the extra activation from the mental concept is added. The words A to F are at their base level of activation. Then, activation from the mental concept is added. The word that then reaches the threshold will be selected for lexical retrieval.

the mental concept has to be zero or less for the lexical entry to be selected.

The base level of activation a word has depends on its number of neighbours: more neighbours means higher base level of activation. Every time one of these neighbours is targeted for lexical selection activation spreads to all other words as well. This frequent activation, even though it is often only partially, results in a higher base level of activation. For now a linear relation between base level of activation *y* and number of neighbours *k* will be assumed, which is specified by some constant *k*. This assumption can be formulated as follows:

[2] $y = kx$

Now, [1] can be rewritten into [3]:

$$T - (kx + \frac{A}{x}) \leq 0$$

$$T - kx - \frac{A}{x} \leq 0$$

$$Tx - kx^2 - A \leq 0$$

[3] $kx^2 - Tx + A \leq 0$

The output of this formula is a parabola, as is shown in Figure 9. Words are available for lexical selection only if the outcome of Formula [3] is above zero. Whether or not the outcome is above zero is determined by the threshold level and the amount of activation coming from the mental concept. For healthy speakers this parabola should always be above zero, which means that all words can be retrieved. Lower processing capacities of aphasic speakers might cause this parabola to appear at a lower position. This would mean that words with an intermediate number of neighbours cannot be retrieved. The result of this would be that Zipf's law for lexical connections shows a gap: no words with intermediate numbers of connections are expected. The size of this gap would then likely correlate with the severity of the aphasia.

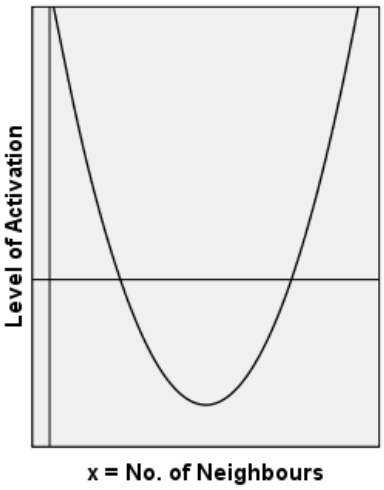


Figure 9. General shape of the outcomes of formula [3]. Here, exact values are irrelevant.

So far it was assumed that *A* is distributed equally over all neighbours. However, it is more likely that the amount of activation per word depends on some word specific coefficient, as in Formula [4].

[4] $Amount\ of\ activation\ per\ word = C\frac{A}{x}$

A schematic representation of such an organisation is given in Figure 10.

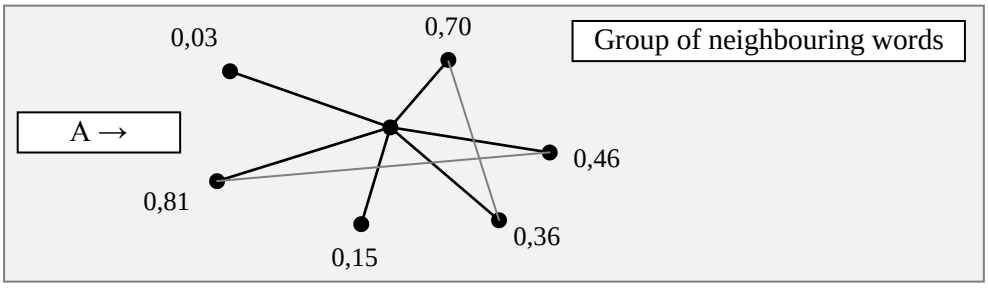


Figure 10. Schematic representation of word specific coefficients in a group of neighbouring words. Note that the coefficients of a group of words do not have to add up to unity.

As discussed in the introduction, a mathematical measure of lexical complexity is entropy: this measure can be used to calculate the degree of complexity of lexical retrieval (De Lange, 2008). Words with higher entropy values require more processing capacity. It can be assumed that C is some function *f* of entropy *H*. The exact relation between *C* and *H* is unknown but has to be positive. It will therefore be referred to as *C* = *f*(*H*) Now, Formula [1] can be rewritten into Formula [5]:

[5] $T - \left(y + f(H)\frac{A}{x}\right) \leq 0$

Combined with Formula [2], Formula [6] can be derived:

$$T - \left(kx + f(H)\frac{A}{x}\right) \leq 0$$

$$T - kx - \frac{f(H) \times A}{x} \leq 0$$

$$kx^2 - Tx - f(H) \times A \leq 0$$

[6] $$kx^2 - Tx + f(H) \times A \geq 0$$

Now, the relation between the number of neighbours and the amount of activation after additional activation from the mental concept depends on entropy. From this it follows that whether or not an aphasic speaker can retrieve a word from his lexicon might depend on the entropy of the word.

What this discussion shows is that the ease of lexical retrieval is likely to depend on the entropy value and on the number of neighbours a word has. The parabolic shape of this dependency might cause words with an intermediate number of connections to be missing from the Zipfian distribution of lexical connections in aphasic speakers. This hypothesis provides a possible explanation for their markedly impaired speech. The question whether or not it is true remains unanswered until it is investigated in future research.

Acknowledgements

This project was joint work with Lizet van Ewijk and will also be incorporated in her dissertation.

References

Afasie Vereniging Nederland. http://www.afasie.nl

Alario, F. –X., Costa, A., & Caramazza, A. (2002). Frequency effects in noun phrase production: Implications for models of lexical access. *Language and Cognitive Processes, 17*, 299-319.

Avrutin, S. (2006). Weak syntax. In: Y. Grodzinsky & K. Amunts (Eds.), *Broca's region*. Oxford: Oxford University Press, 49-62.

Baayen, R. H., Levelt, W. M. J., Schreuder, R., & Ernestus, M. (2007). Paradigmatic structure in speech production. *Chicago Linguistic Society, 43*(The Main Session), 1-29.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht/Boston/London: Kluwer Academic Publishers.

Baayen, R. H. (2007). Storage and computation in the mental lexicon. In G. Jarema, & G. Libben (Eds.), *The mental lexicon: Core perspectives* (pp. 81-104) Elsevier.

Baroni, M. (2008). Distributions in text. In A. Lüdelign, & M. Kytö (Eds.), *Corpus linguistics: An international handbook.* (). Berlin: Mouton de Gruyter.

Burkhardt, P., Avrutin, S., Piñango, M. M., & Ruigendijk, E. (2008). Slower-than-normal syntactic processing in agrammatic Broca's aphasia: Evidence from Dutch. *Journal of Neurolinguistics, 21*, 120-137.

Burkhardt, P., Piñango, M. M., & Wong, K. (2003). The role of the anterior left hemisphere in real-time sentence comprehension: Evidence from split intransitivity. *Brain and Language, 86*, 9-22.

Caplan, D., Waters, G., DeDe, G., Michaud, J., & Reddy, A. (2007). A study of syntactic processing in aphasia I: Behavioral (psycholinguistic) aspects. *Brain and Language, 101*, 103-150.

Corominas-Murtra, B., & Solé, R. V. (2010). Universality of Zipf's law. *Physical Review E, 82*, 011102-1-011102-9.

Cuetos, F., Aguado, G., Izura, C., & Ellis, A. W. (2002). Aphasic naming in Spanish: Predictors and errors. *Brain and Language, 82*, 344-365.

Dahui, W., Menghui, L., & Zengru, D. (2005). True reason for Zipf's law in language. *Physica A, 358*, 545-550.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes, 5*, 313-349.

Ellis, A. W. (2006). Word finding in the damaged brain: Probing Marshall's caveat. *Cortex, 42*, 817-822.

Fellbaum, C. (Ed.). (1998). *Wordnet: An electronic database*. Cambridge, MA: The M.I.T. Press.

Ferrer i Cancho, R., & Elvevåg, B. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE, 3*

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.

Friederici, A. D., Wessels, J. M. I., Emmorey, K., & Bellugi, U. (1992). Sensitivity to inflectional morphology in aphasia: A real-time processing perspective. *Brain and Language, 43*, 747-763.

Gahl, S. (2006). Is frequency a property of phonological forms? Evidence from spontaneous speech.

Grodzinsky, Y. (2000). The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences, 23*, 1-71.

Günther, T., Hofman, M., & Promes, M. (2009). Afasiesyndromen. Twijfels over de klassieke taxonomie. *Logopedie En Foniatrie, 5*, 148-152.

Harley, T. A., & MacAndrew, S. B. G. (2001). Constraints upon word substitution speech errors. *Journal of Psycholinguistic Research, 30*, 395-418.

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.

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.

Laine, M., Tikkala, A., & Juhola, M. (1998). Modelling anomia by the discrete two-stage word production architecture. *Journal of Neurolinguistics, 11*(3), 275-294.

Lange, J. d. (2008). *Article omission in headlines and child language: A processing approach.* Unpublished dissertation.

Laubstein, A. S. (1999). Lemmas and lexemes: The evidence from blends. *Brain and Language, 68*(1-2), 135-143.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 1*(1), 75.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* Mahwah, NJ: Lawrence Erlbaum Associates.

Miller, G. A. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography, 3*, 235-312.

Moscoso del Prado Martín, Fermín, Kostíc, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition, 94*, 1-18.

Nederlandse Taalunie. (2004). *Corpus Gesproken Nederlands.*

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 Altmann* (pp. 545-554). Berlin: Walter de Gruyter.

Popescu, I. -I, Altmann, G., & Köhler, R. (2010). Zipf's law - another view. *Quality and Quantity, 44*, 713-731.

Sahin, N. T., Pinker, S., Cash, S. S., Schomer, D., & Halgren, E. (2009). Sequential processing of lexical, grammatical, and phonological information within Broca's area. *Science, 326*(16 October), 445-449.

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.

Vitevitch, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech, 40*, 211-228.

Zipf, G. K. (1965). *Human behavior and the principle of least effort. an introduction to human ecology*. New York and London: Hafner Publishing company.