Table 4.11 (cont.)

	Syllables per word									
Text #	0	1	2	3	4	5	6	7	8	9
141	3	108	94	84	31	13	1	1	2	
142	1	54	30	32	19	2	1			
143	3	95	52	58	21	7	3			
144	4	86	49	50	22	5	2			
145	5	101	72	90	40	16	4	2		
146	9	307	200	189	90	35	4	2		
147	3	42	23	24	16	9				
148	3	107	73	61	39	19				
149	1	69	36	53	32	7	2			1
150	2	73	49	52	16	9	2			
151	2	52	41	40	27	2				
152	3	46	33	49	20	6	3		2	

5

WITHIN-SENTENCE DISTRIBUTION AND RETENTION OF CONTENT WORDS AND FUNCTION WORDS

August Fenk, Gertraud Fenk-Oczlon

1. Serial Position Effects in the Recall of Sentences

Experiments with free immediate recall of lists of unconnected words usually reveal a saddle-shaped 'serial position curve': high frequency of recall in the items obtaining the first positions ('primacy effect') of the list and those obtaining the last positions ('recency effect'), and, in the words of Murdock Jr., "a horizontal asymptote spanning the primacy and recency effect" (cf. Figure 5.1).

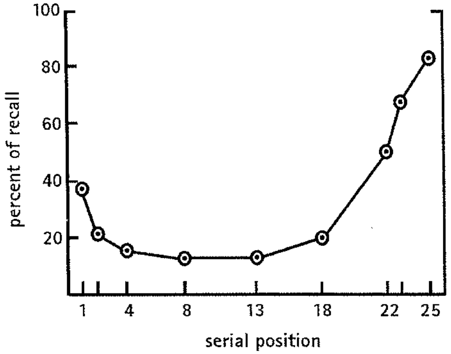

Figure 5.1: A "Typical" Serial Position Curve Resulting From Immediate Free Recall of Unconnected Words (Murdock Jr. 1962: 484, modified)

Murdock Jr. (1962: 488) suggested "that the shape of the curve may well result from proactive and retroactive inhibition effects occurring within the list itself." Assumptions regarding the underlying mechanisms became more differentiated

P. Gryzbek (ed.), Contributing in the Science of Text and Language, 157-170.

in later experiments introducing – in addition to input-order – sense-modality of list presentation as a second independent variable (Murdock & Walker 1969) and evaluating – in addition to frequency of recall – output-order in recall as a second dependent variable (Fenk 1979). Many of the relevant psychological findings seem to be interesting for linguistics as well: the results reported e.g. by Murdock Jr. (1962) suggesting that the recency effect extends over the last 7 plus minus 2 words; the observation of Murdock & Walker (1969) that in auditory presentation the recency effect is higher and more extensive than in visual presentation ('modality effect'); the observation that 'sequential clusters', i.e. word groups with the words recalled in the same order as represented, are in the recency part of auditorily presented word strings significantly larger and significantly more frequent than in the recency part of visually presented word strings – in series of unconnected nouns as well as in real sentences, and despite the tendency to start the recall of auditorily presented word strings with words and word groups from the end of the string (Fenk 1979: 14). Of particular linguistic interest is the question whether the serial position effects shown in the recall of lists of unconnected words show in the recall of real sentences as well. Are the underlying processes also efficient in real sentence processing and in connected discourse?

Indications reported so far (Jarvella 1971; Fenk 1979, 1981; Rummler 2003) are not fully convincing: In Jarvella's study, subjects were presented sentences like "...Having failed to disprove the charges, Taylor was later fired by the president" (p. 410). The fragment starting with "Taylor..." is not only localized in the 'recency part' of the sentence but also represents the main constituent of this sentence, so that one has to suspect a confusion of the effects of these conditions. The marked and plateau-shaped 'recency effect' in the serial position curves (p. 411) might suggest some additive effects on the performance of recall (see Figure 5.2).

Rummler (2003: 96) states that the so-called "subordination effect" (subordinating conjunctions achieve a better verbatim recall than coordinating conjunctions) mainly comes off by better retention of the second clause. This better recall of the second clause might again be a consequence of the restricted "span" of the recency effect and/or of the fact that in the subordinating conjunction the internal redundancy is higher and the informational content (and cognitive load) of the second clause lower than in the coordinating conjunction.

The serial position curve reported by Fenk (1981: 25) shows a marked recency effect (only) in auditory presentation of the sentence. (And, in addition, an 'inverse modality effect', i.e. a superiority of visual presentation in the primacy part.) But these results originate from only two different sentences presented simultaneously in two different sense modalities.

For a more systematic approach, the subjects in an experiment by Auer, Bačik & Fenk (2001) were presented a text by Glasersfeld (1998). Speech was interrupted after some of the sentences by a certain signal. Subjects were instructed to write down as much as they could remember from the last sentence before the test pause. Since the position of a word is by far not the only factor determining its chance to be recalled, the serial position curves obtained differ not only from the "ideal" curve (Figure 5.1) that approximately shows when different groups of subjects are presented different lists with the single items changing their positions in systematic rotation. They differed also from sentence to sentence, since these individual sentences ($n = 10$) differed in all possible respects – lexemes, syntactic structure, length. Nevertheless the family of curves shows a (rather weak) primacy effect and a marked recency effect. Data from this experiment were re-analyzed in order to investigate further questions.

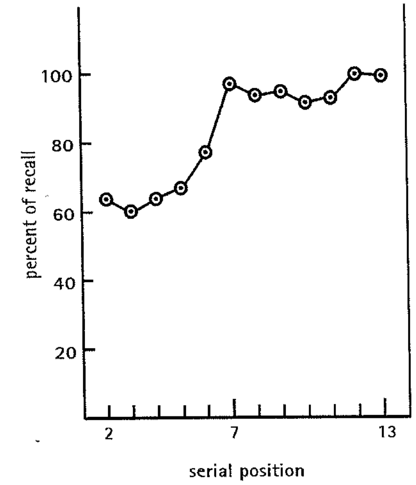

Figure 5.2: One of the Serial Position Curves For Free Recall of Sentences (Jarvella 1971: 414, modified)

2. Wordclass-specific Effects on the Serial Position Curve?

2.1 Two hypotheses

The aim of our statistical reanalysis was to test the following assumptions:

Hypothesis 1
In content words the likelihood of recalling is higher than in function words.

Hypothesis 2

In the recency part the difference predicted in assumption I will be smaller than in earlier parts of the serial position curve.

The basic consideration leading to these assumptions was that the fundamental difference between content words and function words (a) should be in some way reflected in sentence processing, especially in our 'semantic memory' (b).

Ad (a):

Within linguistics, content words are often characterized as 'open class words' opposed to function words as 'closed class words'. Rather relevant with respect to our hypotheses is the fact that content words are more significant for the specific topic, just for the particular and concrete 'content' of texts and sentences, while the function of function words is to bring about certain references between more or less exchangeable contents and content words. In brief: The relevant division here is between context-specific content words and rather context-independent function words. Function words are, moreover, rather short, very frequent, and 'multifunctional' in the sense of Zipf (1949).

Ad (b):

A widely accepted model concerning our memory says: After having extracted the meaning of an actual clause, its verbatim form (words and syntax) is rapidly lost from memory, while the meaning is preserved (and affects e.g. our interpretation of the following clauses). This conception is strongly influenced by Sachs (1967). Her findings could be reproduced in later experiments by Luther & Fenk (1984) which showed moreover that this outcome is not grounded in the nature and incapability of our long term memory but is the result of a cognitive strategy which is successful under 'normal' conditions, i.e. in the absence of the instruction and motivation to concentrate on other aspects of sentences.

This principle – rapid loss of the form after the meaning has been extracted – is actually also supported by two findings already mentioned in the present paper:

(i) The tendency to repeat some of the words in the 'input order' especially from the recency part of auditorily presented word strings (Fenk 1979: 14) indicates that verbatim representation is a speciality of immediate acoustic memory.

(ii) The plateau at the end of Jarvella's (1971) serial position curves. Jarvella's comment on his findings:

> "Various verbatim measures of recall support only the immediate sentence and immediately heard clause as retrievable units in memory" (p. 409). "Apparently only these immediate sentences hold a retrievable form in memory; this form also leads to superior recall of their most recent clause. On the other hand, recall of previous sentences indicates that they had received a relatively thorough semantic interpretation. It appears that the propositional meaning of sentences was remembered shortly after they were heard, although, as measured by verbatim recall, the form of sentences was quickly forgotten" (p. 415).

2.2 Method

In order to test our assumptions, each of the test-sentences used in the Auer, Bačik & Fenk experiment was divided into four quarters, rounding off where necessary. The first quarter (I) was defined as the primacy part of the sentence, II and III taken together as the medium part, and the last 25 percent of the words (IV) as the recency part. Then we determined, separately for the three parts (I, II+III, IV) of each sentence, the number of content words – nouns, verbs, adjectives, manner adverbs – and the number of function words such as articles, prepositions, pronouns, conjunctions, negations, particles, auxiliary verbs.

Our operationalization of 'primacy part' (first 25% of words) and 'recency part' (last 25% of words) might, at a first glance, appear as a rather arbitrary and rough method, since a 'quarter' means different things in sentences of differing length: e.g. five words in a 20-word sentence or ten words in a 40-word sentence. But the alternative – to define the primacy part and the recency part in terms of a fixed number of words – would again be arbitrary: How many words should be fixed? And it would restrict the application to rather long sentences: A fixed number of, let us say, six words for the primacy part and six words for the recency part would, in the case of a 12-word sentence, reduce the 'medium part' to zero, and would exclude shorter sentences altogether.

Our operationalization, however, offers a wide range of applications and establishes a firm proportion between, on the one hand, the primacy and recency part, and, on the other hand, the part in between and the sentence as a whole. And it has proved to bring about significant results.

2.3 Results

A problem for the quantification of a primacy and recency effect in our two word classes was the unexpected observation that the proportion of content words to function words showed a considerable variation between the interesting parts of the sentences. Thus, a quantification in absolute terms did not make much sense, and the recall scores had to be related to the number of words presented. Table 5.1 lists the results – number of words occurring, number of words recalled, and the 'relative' recall scores R/O.

Table 5.1: Number of Words Occurring (O) in and Recalled (R) from the First Quarter (I), the Medium Part (II + III), and the Last Quarter (IV) of a Total of 10 Sentences; R/O = Mean Frequency of Recall Per Word Given

	I			II + III			IV		
Wordclasses	O	R	R/O	O	R	R/O	O	R	R/O
Content Words (C)	22	55	2.5000	75	152	2.0267	37	132	3.5676
Function Words (F)	37	68	1.8378	68	98	1.4412	25	75	3.0000
Total ($C + F$)	59	123	2.0847	143	250	1.7483	62	207	3.3387
Difference ($C - F$)			0.6622			0.5855			0.5676
Level of significance (p)			< 5%			< 5%			< 1%

The results of the statistical evaluation (Wilcoxon tests) in words:

- The primacy effect – the gradient we can see in Figure 5.3 between middle part and primacy part – was not significant.
- The recency effect, i.e. the gradient between middle part and recency part, was more marked (Figure 5.3) and was significant in all possible evaluations: in the content words ($p < 1\%$), in the function words ($p < 2\%$), and when both word classes were taken together ($p < 1\%$).
- **Hypothesis 1** was clearly confirmed: Level of relative recall scores was significantly higher in content words than in function words throughout the sentences. (Table 5.1 presents in its lowest line the error probabilities for parts I, II+III, and IV).
- **Hypothesis 2** would predict a convergence between the recall curves for content words and function words at least between the middle part and the recency part (Figure 5.3). Actually there is, as can be seen from the values in Table 5.1, a slight convergence from I to II+III and from II+III to IV. But in both cases this convergence is far from significant.

3. Three More or Less Hypothetical Regularities

The formulation of the first of the following assumptions is motivated by the occasional observation that our test sentences taken from a Glasersfeld text showed a tendency of an increase of content words and a decrease of function words during a sentence. As to this tendency we carried out a little follow-up study (3.1). Results strongly indicate that this is a general tendency at least in German texts. And if our tentative explanation (section 4) of this regularity holds, its scope should not be restricted to German texts.

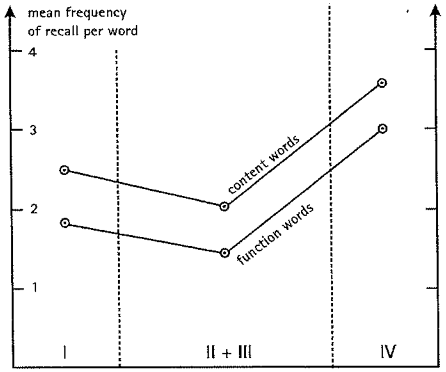

Figure 5.3: Differences Between Content and Function Words in a Serial Position Curve For Free Recall of Sentences of Different Length (I = first quarter, IV = last quarter)

Regularities 3.2 and 3.3 have the status of as yet unexamined lawlike hypotheses.

Regularity 3.2 proceeds from the assumption that the token frequency of function words is higher than the token frequency of content words. If these function words tend to occupy initial positions of sentences, this should contribute to the regularity "the more frequent before the less frequent". This statistical regularity has proven to be the most powerful one in the explanation of word order in frozen conjoined expressions (Fenk-Oczlon 1989), and it seems that its range of validity can be extended on clauses in general. In this present paper we will state this generalized rule mainly as an inferential step to our third regularity (3.3) which perfectly fits with the topic of this volume on "word length".

3.1 Decrease of Function Words and Increase of Content Words Within Sentences

As Table 5.1 shows, function words tend to decrease and content words to increase in the course of a sentence. Figure 5.4 illustrates these tendencies. Despite the small sample of only ten sentences, the relevant differences proved to be significant in the Wilcoxon test (Table 5.2).

These differences in the distribution of the instances of the two word classes were, as already mentioned in section 2.3, a problem for a simple evaluation of the recall scores. But we suspected that it might indicate an interesting

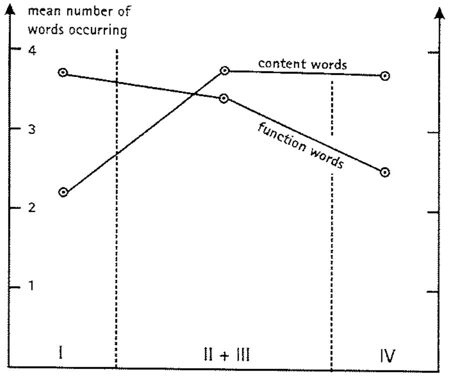

Figure 5.4: Differences Found in the Within-Sentence Distribution of Content and Function Words

phenomenon within the scope of quantitative linguistics – provided, that these tendencies are not a special feature of a certain text of a certain author. A pilot study was conducted in order to find some indications of possible generalizations of this tendency. The sample of authors was increased – nine more German text passages, four of them from scientific books, five from literary books. Taken together with the already analysed text passages from Glasersfeld this is a sample of ten (five scientific, five literary) text passages, and a sample of ten sentences from each of these passages, i.e. a total of $10 \times 10 = 100$ sentences. (Source texts are listed at the end of the paper.)

Table 5.2: Differences Between the Frequency of Function Words and Content Words Occurring in the Primacy Part (I) and in the Recency Part (IV) of 10 Sentences from a Text by Glasersfeld (1998)

difference	quarter I – quarter IV	error probability
function words	12	$p < 1\%$
content words	−15	$p < 1\%$
difference	function w. – content w.	
quarter I	15	$p < 1\%$
quarter IV	−12	$p < 5\%$

The evaluation was carried out by a student who did not know our assumptions. She was instructed not to collect ten successive sentences from each passage into the sample, but – where possible – each third sentence. Sometimes she had to overleap more than two sentences, e.g. when one of the intervening ‘sentences’ was too short (less than four words) or the heading of a new section. As already suggested by Niehaus (1997: 221), a colon was accepted as the end of the sentence when the following word started with a capital letter. The results: The gradient of the decrease of function words and increase of content words from the first quarter to the last quarter is not as steep as in Glasersfeld, but still significant (Fenk-Oczlon & Fenk 2002). These results suggest that the tendency of function words to decrease and of content words to increase in the course of a sentence is a general tendency at least in German texts. And the provisional results of Müller (in preparation) indicate that this tendency is not restricted to German texts.

3.2 The More Frequent Before the Less Frequent

This regularity was originally stated in order to explain and predict the order of the two main elements forming a frozen conjoined expression such as *knife and fork*, *salt and pepper*, *peak and valley*. From all the rules examined (e.g. “short before long”, “the first word has fewer initial consonants than the second”), the rule “the more frequent before the less frequent” showed the highest explanatory power as to the word order in 400 freezes (Fenk-Oczlon 1989). Our regularity 3.1 should establish or enhance such a tendency in sentences as well to the effect that regularity 3.2 is not restricted to freezes.

If the tendency “the more frequent before the less frequent” fits to sentences as well – as a consequence of 3.1 or for whatever reason – then it is plausible to assume a further regularity:

3.3 Increase of Word Length in the Course of a Sentence

A more general regularity of this sort was already postulated in Behaghel (1909): “Das Gesetz der wachsenden Glieder”, i.e. “the law of increasing elements, parts, links, constituents. . . ”. Behaghel illustrates this law with many examples from classical texts in a variety of languages such as ancient Greek, Latin, Old High German and German. In most of his examples the comparison was between word groups of different size or between single words and word groups: *auf der **Türbank** und im **dunklen Gang*** (p. 110), ***ih** inti **father min*** (p. 111). In a little experiment by Behaghel the subjects got four sheets of paper with the following words and word groups: *Gold / edles Geschmeide / und / sie besitzt*. They were instructed to form a sentence from these fragments, and the result was always the same: sie besitzt ***Gold*** und ***edles Geschmeide***. (Behaghel 1909: 137). He offers the following interpretation:

> Man wird nicht nur die länger dauernde Arbeit auf den Zeitraum verlegen, wo man den Abschluß leichter hinausschieben kann; man wird auch, wenn man sich Zeit lassen kann, die Arbeit gründlicher tun, mehr ins Einzelne gehen, oder, sprachlich ausgedrückt: man wird nicht nur für den umfangreicheren Ausdruck die spätere Stelle wählen, sondern auch für die spätere Stelle den umfangreicheren Ausdruck sich zubereiten. So bildet sich unbewußt in den Sprachen ein eigenartiges rhythmisches Gefühl, die Neigung, vom kürzeren zum längeren Glied überzugehen; so entwickelt sich das, was ich, um einen ganz knappen Ausdruck zu gewinnen, als das *Gesetz der wachsenden Glieder* bezeichnen möchte. (Behaghel 1909: 138f.)

Our regularity "increase of word length during a sentence" is similar to Behaghel's law but more specific in that it is localized at the single word level. A special case of Behaghel's law, so to speak! At present we cannot offer results of empirical tests of this lawlike assumption. (But see Müller, in preparation.) But we can contribute two new perspectives:

1. An operationalization that allows for a statistical examination of the law: Define words as the relevant 'Glieder' or 'constituents', determine their 'size' in terms of 'number of syllables', and compare the mean size of words in the early and late parts of sentences.
2. An interpretation specifying a concrete factor that might at least contribute to the rhythmic pattern described by Behaghel. This factor is the concentration or accumulation of function words in the first parts of clauses (sentences, subordinate clauses). And since function words are generally extremely frequent and frequent words tend – for economic reasons – to be rather short (Zipf 1929, 1949), the concentration of these rather short units in the first part of clauses results in an increase of the mean word length in the course of a sentence. This hypothesized tendency will of course depend on the respective language type and is expected to be more pronounced in languages with a tendency to agglutinative morphology and a tendency to OV order.

4. Conclusions

A possible explanation for our regularity "decrease of function words and increase of content words": In a running text, almost any clause has to refer to what was said in the clauses before ("old before new", "topic before comment", "theme before rheme"). This reference is – most probably not only in German texts – first of all brought about by function words (e.g. anaphoric pronouns, conjunctions) right at the beginning of the new clause. If this is an appropriate explanation of our regularity 3.1, then it is – indirectly – also relevant for the hypothesized regularities 3.2 "the more frequent before the less frequent" and 3.3 "within-sentence increase of word length". The last steps of the arguments in other words: Function words accumulate at the beginning of clauses; they are very frequent and 'therefore' very short in terms of number of syllables; members of our second word class 'content words' are, on average, composed

of a higher number of syllables, and the number of these content words tends to increase during the sentence. This means, first of all: The regularity "the more frequent before the less frequent" found in frozen binomials holds true for sentences as well. As a consequence, one may expect an increase of word length in the course of a sentence. All the regularities outlined above ("the more frequent before the less frequent", "short before long") fit to and contribute to the more general law (Fenk-Oczlon 2001) of an economic and rather 'constant' flow of linguistic information.

5. Appendix: Sources of the Test Sentences

Bachmann, I. (1980 [1971]). Malina. Frankfurt: Suhrkamp Verlag (suhrkamp taschenbuch 641).
▷ pp. 200–202, beginning with section "Malina ist . . . "

Frisch, M. (1975). Montauk. Gütersloh: Bertelsmann Reinhard Mohn OHG.
▷ pp. 157–159, beginning with section "Money"

Gigerenzer, G., Swijtink, Z., Porter, Th., Daston, L., Beatty, J., Krüger, L. (1999). Das Reich des Zufalls. Heidelberg/Berlin: Spektrum Akademischer Verlag.
▷ pp. 212–213, beginning with section "5.8 Diskontinuität, eine Grundlage aller Veränderung"

Glasersfeld, E. von (1998). Konstruktivismus statt Erkenntnistheorie. In: W. Dörfler & J. Mitterer (eds.), Ernst von Glasersfeld – Konstruktivismus statt Erkenntnistheorie. Klagenfurt: Drava Verlag.
▷ pp. 11–17

Hesse, H. (1972). Der Stéppenwolf. Gütersloh: Bertelsmann Reinhard Mohn OHG
▷ pp. 269–271, beginning with section "Die Fremdenstadt im Süden"

Mann, Th. (5. Aufl. 1997 [1947]). Doktor Faustus. Frankfurt a. M.: Fischer Taschenbuch Verlag.
▷ pp. 47–50, beginning with section "VI"

Musil, R. (1960; A. Frisé, ed.). Der Mann ohne Eigenschaften. Stuttgart: Deutscher Bücherbund.
▷ pp. 445–447, beginning with section "98. Aus einem Staat, der an einem Sprachfehler zugrundegegangen ist"

Niehaus, B. (1997). Untersuchung zur Satzlängenhäufigkeit im Deutschen. In: Best, K.-H. (ed.), The Distribution of Word and Sentence Length. Glottometrika 16, Quantitative Linguistics 58, 213–275. Trier: WVT Wissenschaftlicher Verlag Trier.
▷ pp. 263–264, beginning with section "6. Ausblick"

Spies, M. (1993). Unsicheres Wissen. Heidelberg/Berlin/Oxford: Spektrum Akademischer Verlag.
▷ pp. 20–21, beginning with section "3. Perspektive: Kognitive Modelle der Informationsverarbeitung"

Stegmüller, W. (1957). Das Wahrheitsproblem und die Idee der Semantik. Wien: Springer-Verlag.
▷ pp. 38–40, beginning with section "III. Die Trennung von Objekt- und Metasprache als Weg zur Lösung und die Idee der Semantik als exakter Wissenschaft. Semantische Systeme von elementarer Struktur"

References

Auer, L.; Bačik, I.; Fenk, A.
2001 "Die serielle Positionskurve beim Behalten echter Sätze." Vortrag am 26.10.2001 im Rahmen der 29. Österreichischen Linguistiktagung in Klagenfurt.

Behaghel, O.
1909 "Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern", in: *Indogermanische Forschungen, 25*; 110–142.

Fenk, A.
1979 "Positionseffekte und Reihenfolge der Wiedergabe bei optisch und akustisch gebotenen Wortketten", in: *Archiv für Psychologie / Archives of Psychology, 132(1)*; 1–18.

Fenk, A.
1981 " 'Ein Bild sagt mehr als tausend Worte...?' Lernleistungsunterschiede bei optischer, akustischer und optisch-akustischer Präsentation von Lehrmaterial", in: *AV-Forschung, 23*; 3–50.

Fenk-Oczlon, G.
1989 "Word frequency and word order in freezes", in: *Linguistics, 27*; 517–556.

Fenk-Oczlon, G.
2001 "Familiarity, information flow, and linguistic form." In: Bybee, J.; Hopper, P. (eds.), *Frequency and the emergence of linguistic structure.* Amsterdam / Philadelphia. (431–448).

Fenk-Oczlon, G.; Fenk, A.
2002 "Zipf's Tool Analogy and word order", in: *Glottometrics, 5*; 22–28.

Jarvella, R.J.
1971 "Syntactic processing of connected speech", in: *Journal of Verbal Learning and Verbal Behavior, 10*; 409–416.

Luther, P.; Fenk, A.
1984 "Wird der Wortlaut von Sätzen zwangsläufig schneller vergessen als ihr Inhalt?", in: *Zeitschrift für experimentelle und angewandte Psychologie, 31*; 101–123.

Müller, B.
in prep. *Die statistische Verteilung von Wortklassen und Wortlängen in lateinischen, italienischen und französischen und italienischen Sätzen.* Phil. Diss., University of Klagenfurt.

Murdock, B.B., Jr.
1962 "The serial position effect in free recall", in: *Journal of Experimental Psychology, 64*, 482–488.

Murdock, B.B.; Walker, K.D.
1969 "Modality effects in free recall", in: *Journal of Verbal Learning and Verbal Behavior, 8*; 665–676.

Niehaus, B.
1997 "Untersuchung zur Satzlängenhäufigkeit im Deutschen." In: Best, K.-H. (ed.), *The Distribution of Word and Sentence Length.* Trier. (213–275). [= Glottometrika 16, Quantitative Linguistics; 58]

Rummler, R.
2003 "Das kurzfristige Behalten von Sätzen", in: *Psychologische Rundschau, 54(2)*; 93–102.

Sachs, J.S.
1967 "Recognition memory for syntactic and semantic aspects of connected discourse", in: *Perception & Psychophysics, 2(9)*; 437–442.

Zipf, G.K.
1929 "Relative frequency as a determinant of phonetic change", in: *Harvard Studies in Classical Philology, 40*; 1–95.

Zipf, G.K.
1949 *Human behavior and the principle of least effort. An introduction to human ecology.* Cambridge, Mass. [^{2}1972, New York.]

6

ON TEXT CORPORA, WORD LENGTHS, AND WORD FREQUENCIES IN SLOVENIAN

Primož Jakopin

1. Introduction

From the first beginnings in the mid-1990s, availability of electronic text corpora in Slovenian, all with an Internet user interface, has grown to a level comparable to many European languages with a long history of quantitative linguistic research. There are two established corpora with 100 million running words, an academic one which is freely accessible and a commercial one, prepared by industrial and academic partners. The two are complemented by a sizeable collection of works of fiction, available for reading in a free virtual library and several specialized corpora, compiled for the needs of particular institutions. The majority of Slovenian newspapers are also accessible online, at least in the form of selected articles.

Lists of word forms with frequencies can be downloaded in chunks of 1000 from the *Nova beseda* corpus, and a lemmatization service is also available from the companion page (`http://bos.zrc-sazu.si/dol_lem.html`). Online translation from Slovenian into English for short texts (up to 500 characters) is already at hand (`http://presis.amebis.si/prevajanje`), with English-Slovenian in preparation.

The basic infrastructure for word-length analysis is in place and in the following chapters these topics are discussed in some more detail.

2. Online Text Corpora

There are two online text corpora in the narrow sense of this word, each 100 million running words in size and each equipped with an Internet user interface including a concordancer and some other searching facilities. Other text collections have been built with different uses in mind and they complement the Slovenian corpus scene.

P. Gryzbek (ed.), Contributing in the Science of Text and Language, 171-185.