Acta Linguistica Hungarica, Vol. 49 (3–4), pp. 385–405 (2002)

MULTILINGUAL STATISTICAL TEXT ANALYSIS, ZIPF'S LAW AND HUNGARIAN SPEECH GENERATION*

GÉZA NÉMETH – CSABA ZAINKÓ

Abstract

The practical challenge of creating a Hungarian e-mail reader has initiated our work on statistical text analysis. The starting point was statistical analysis for automatic discrimination of the language of texts. Later it was extended to automatic re-generation of diacritic signs and more detailed language structure analysis. A parallel study of three different languages—Hungarian, German and English—using text corpora of a similar size gives a possibility for the exploration of both similarities and differences. Corpora of publicly available Internet sources were used. The corpus size was the same (approximately 20 Mbytes, 2.5–3.5 million word forms) for all languages. Besides traditional corpus coverage, word length and occurrence statistics, some new features about prosodic boundaries (sentence initial and final positions, preceding and following a comma) were also computed. Among others, it was found that the coverage of corpora by the most frequent words follows a parallel logarithmic rule for all languages in the 40–85% coverage range, known as Zipf's law in linguistics. The functions are much nearer for English and German than for Hungarian. Further conclusions are also drawn. The language detection and diacritic regeneration applications are discussed in detail with implications on Hungarian speech generation. Diverse further application domains, such as predictive text input, word hyphenation, language modelling in speech recognition, corpus-based speech synthesis, etc. are also foreseen.

1. Introduction

As language and speech technology applications gain an increasingly widespread use in several languages/countries, it is important to re-examine the issue of how much difference exists between English (in most cases the first language for both technologies and applications) and other languages. These differences are studied and described in detail in linguistics but they are rarely quantified and used by technology developers.

* The authors are thankful for the help of Manuel Kaesz in collecting the text corpora of equal size for German and English. The current paper is an updated and extended version of the work described in Németh – Zainkó (2001). This research was partly supported by the Pannon GSM Professorship scheme for the first author and by a PhD student grant of Timber Hill Ltd for the second author.

In this paper a parallel study of three linguistically different languages—Hungarian, German and English—will be described, using text corpora of a similar size of standard texts and different versions of the Bible. Besides traditional corpus coverage, occurrence statistics, weighted and unweighted word length, some new display properties, features about prosodic boundaries (sentence initial and final positions, preceding and following a comma) were also computed. Examples of applying the above-mentioned results in practical applications will also be given.

2. Text corpora

Corpora of publicly available Internet sources was used. Word units are defined as characters between white spaces. It is important to note here that inflected forms of the same root count several times according to this definition. In order to avoid distortions, we tried to filter out asterisks, dashes, slashes, round and square brackets, and other non-relevant characters from corpora. We could not drop all non-real-word strings, because sentence length computations would have been seriously affected. Most of the non-word strings retained are numbers, Roman numbers and abbreviations.

The corpus size of standard texts was the same (approximately 20 Mbytes, 2.5–3.5 million word forms) for all languages. The Hungarian corpus was selected from texts larger than 50 kbytes in the Hungarian Electronic Library (HEL, approximately 2.5 million words). The German corpus was collected from similar material of the Gutenberg project (approximately 3.1 million words). The English corpus was collected from English sections of HEL (approximately 3.5 million words). All corpora contain various texts (literature, newspaper, etc.). The similar size of corpora was a major factor during collection as we wanted to avoid distortions among languages caused by greatly differing coverage and topic domains. For the purpose of comparison, electronic versions of the Bible in Hungarian, German and English were also studied (King James Bible, American Standard Version of the Bible, Elberfelder Bible, Katolikus Biblia).

In order to compare coverage effects, a larger corpus of approximately 80 million words (denoted by Hungarian2) was generated for Hungarian by adding data to the HEL corpus from online newspapers and the Digital Literary Academy (13 million word forms) and combining it with a list of 700,000 words which was derived from up-to-date texts containing 21 million words (Hungarian National Corpus, see Váradi 9). Hungarian2 (80 million word

forms) contains approximately 2 million different words. We have also processed derived data from the British National Corpus (BNC, 89 million word forms, see Kilgarriff 2) filtered the same way as the other corpora.

3. Statistical analysis

3.1. Corpus coverage

Looking at both theoretical studies and practical applications in speech recognition, it seems as if a 20,000 word vocabulary had some magic feature because it is a very frequently used number (sometimes together with language difference warnings, e.g., Gibbon et al. 1, 41–5; Roukos 7). Our results confirm this feature **for English**. Looking at Table 1, it can be seen that such a vocabulary yields a 2.5% theoretical minimum error rate, which coincides with results of other studies. It is important to note, however, that in order to reach the same error rate limit, **German** requires a vocabulary **4 times as large** and it grows by **20 times for Hungarian**.

Table 1

Number of most frequent words required by corpus coverage

LANGUAGE	CORPUS COVERAGE		
	75%	90%	97.5%
English	1,250	5,800	20,100
German	2,000	14,550	80,000
Hungarian	10,650	70,000	400,000

Table 2 gives the coverage rate using the 1,000, 20,000 and 100,000 most frequently occurring words in the vocabulary. One reason for the appearance of 20,000 word systems for non-English Western European languages might be that similarly to German, they reach above 90% coverage, which can be acceptable in some cases. It is clear however that an 80% coverage rate is not acceptable in most applications. It is probable that for highly inflecting languages (Hungarian, Finnish and Slavic languages) far larger vocabularies are to be applied if similar processing methods are used as in English. The above 70% coverage of 1,000 words in English might be an explanation why many English teachers claim (at least in Hungary) that flexible and quick use of such a vocabulary is enough for everyday communication in most situations. The same argument may be valid for other quick learning techniques as well.

Table 2
Some examples of corpus coverage

LANGUAGE	NUMBER OF MOST FREQUENT WORDS		
	1,000	20,000	100,000
English	72.8%	97.5%	(100%)
German	69.1%	91.8%	98.1%
Hungarian	51.8%	80.7%	92.0%

It is a popular tool in computational linguistics to use the frequency-rank distribution plot of text corpora. According to Zipf's law it is supposed that such a plot follows rule (1), where C is a normalising constant and b is around 1.

(1) $freq(rank) = C * rank^{-b}$

Another approximate equation is (2) from Lavalette (see Popescu 6):

(2) $freq(rank) = C * ((rank * maxrank)/(maxrank - rank + 1))^{-b}$

This is better in the range of low frequency items than the original Zipf's law. A comprehensive bibliography of this problem can be found in Li (3). Figure 1 shows the results obtained for our standard text corpora. It can be clearly seen that in the 10–10,000 range there is a close coincidence of English and German, while the slope of the Hungarian curves (which run nearly parallel) is slightly different from both other corpora. The upper and lower regions of all corpora seem to be rather irregular. The slope of the BNC corpus is very similar to that of the smaller Hungarian corpus. The limitations of the original Zipf's law and the Lavalette law are illustrated by fitting them to the similarly large corpora of Hungarian2 and the BNC. It seems that proposed measures for extending Zipf's rule to the whole range are not successful for the corpora we studied.

From a practical point of view, we consider the coverage-rank distribution plot far more useful than the frequency-rank distribution. It is essentially the integral of the ranking plot and normalised to 1. Our results are given in Figure 2. The vertical axis is linear while the horizontal one is logarithmic in order to ensure a display ratio of 1 to 10,000,000. It is an interesting result that in the above 40% range all relatively large corpora (except Hungarian) result in parallel lines. The functions in the 40–85% range could be well approximated by straight lines. The relationship is nearly purely exponential

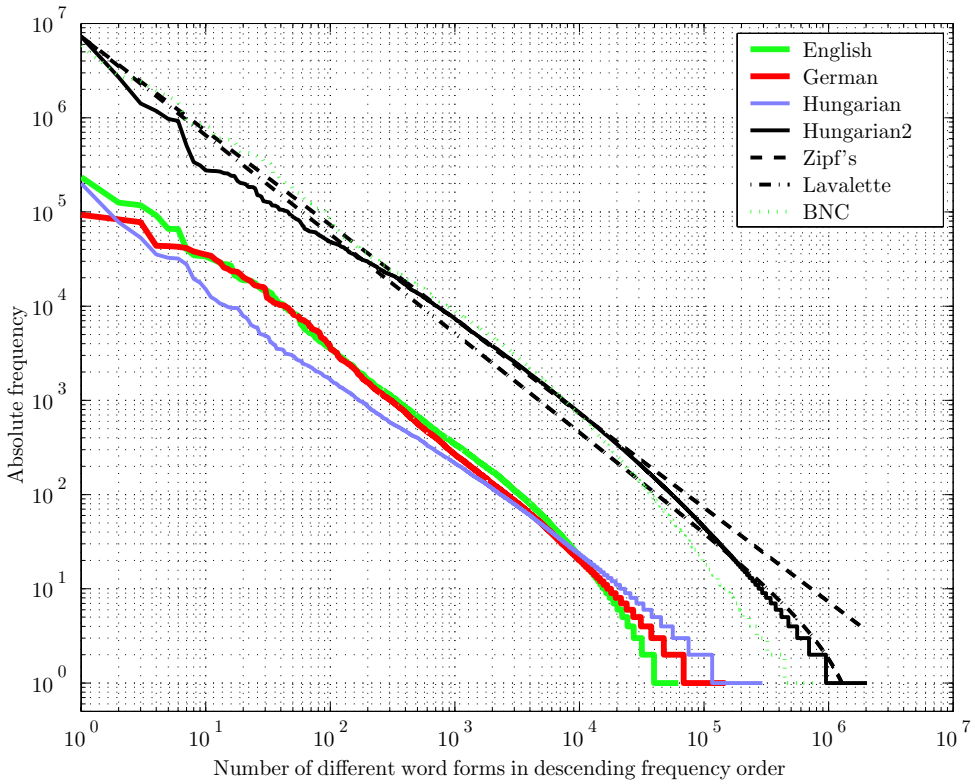

Fig. 1
Ranking of standard texts

and follows the Zipfian rule. It is important to note that this is true for the middle range only and that the critical lower and higher ranges are different.

It seems that the Hungarian2, German and English corpora display similar properties as they run parallel above 40%. The German line runs much nearer to the English one than to the Hungarian as expected according to theoretical assumptions. The English corpus differs in coverage by approximately a factor of 2 from the BNC coverage line over 40%, the shape being very similar. The Hungarian corpus seems to be too small to give even approximative results above 95%. It is also clear from the figure that above 95% there is a saturation effect, i.e., disproportionally large number of new words are needed for a small increase in coverage (e.g., for Hungarian2 by approximately doubling the vocabulary—43,000 to 90,000—one can jump from 85%

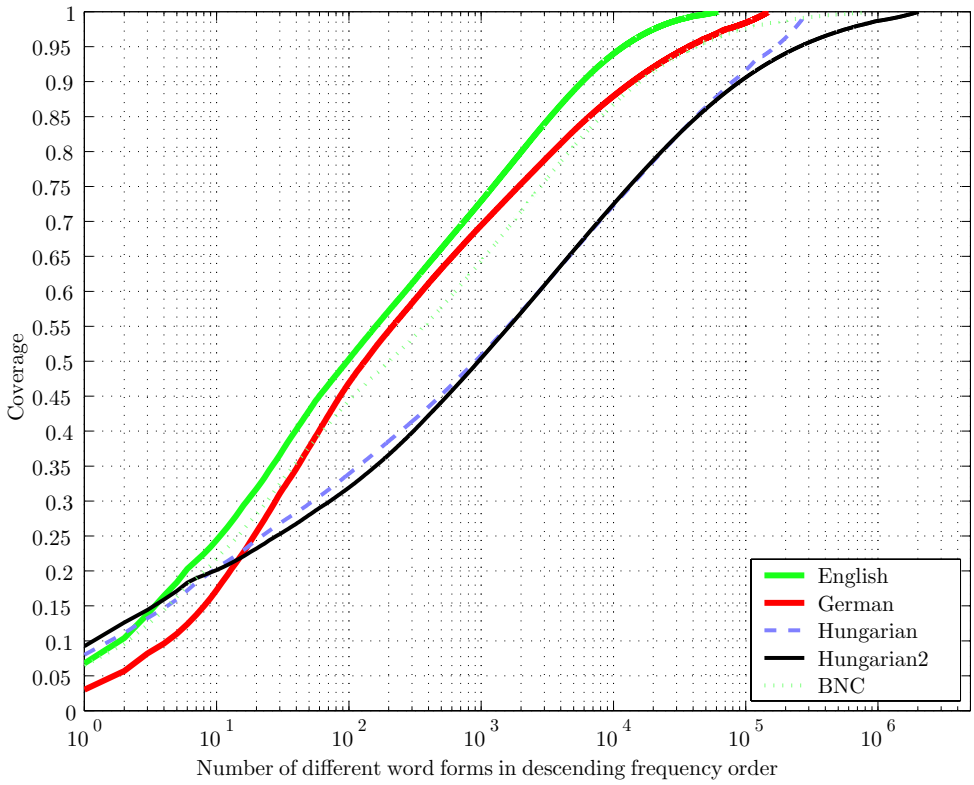

Fig. 2
Corpora coverage by the most frequent words
(logarithmic horizontal scale) of standard texts

to 90%, but increasing it from 254,000 to 470,000 raises coverage from 95% to 97% only). Even this section could be well approximated by straight lines on the figure. It may be the case that above 97% Zipf's law could also be applied with a different b constant than in the 40–85% range.

Although three languages are not enough for making generic statements for several languages, it is worth mentioning that the shape of the coverage functions is surprisingly similar above 40%. It might be worthwhile to conduct similar studies for several languages. If the functions are similar, a single measure for comparing language complexity in case of corpora of similar size might be used. We propose a measure of the number of words needed for covering 95% of a sufficiently large, representative corpus of a language. The

corpus should be regarded as representative if it reaches the saturation state above 95%. The name of the measure could be COV95. Our COV95 measures for approximately 3 million word corpora (the exact corpus size is the first, the name of the language is the second parameter) are as follows:

(3) COV95 (3.5M, English) = 11,859
COV95 (3.1M, German) = 36,982
COV95 (2.5M, Hungarian) = 168,510

It is also worth looking at the lower end of the figure. The 10 most frequent words cover 15–25%, the first 100 cover 35–50%, while the first 1000 provide 50–75% coverage. This means that in several cases (e.g., diacritic regeneration, speech synthesis, language and keyword detection) careful handling of relatively few words can provide significant improvements.

Figure 3 (overleaf) illustrates a very problematic aspect of corpus based approaches. It is clear that even for English, which contained only 62,000 different word forms in a 3.5 million corpus, nearly 40% of the 62,000 different units (at least 20,000 words) appeared only once in the corpus. So even if one collects a huge corpus for training a system, in case of a real-life application there is a very great probability that quite a few new items (related to the training corpus) will appear. If the corpus is large enough—such as the BNC for English—a very large ratio of rare items will appear only once. For Hungarian the problem is even harder. In a practically convincing case one should collect either such a big corpus that all items should fall in the rightmost column (i.e., appearing at least five times in the corpus) or apply rule-based approaches. Often the combination of both techniques may provide the best solutions.

It is important to note that, although the Hungarian corpora had far more word forms than the German one did, this distribution is very similar for both languages.

3.2. Comparative results for English, German and Hungarian Bible versions

In the closed topic domain of the Bible the similarities of English and German have been demonstrated in the frequency-rank plot in the 10–1,500 range. Note, however that even these plots are rather different outside that range. Hungarian displayed largely different properties. The two different English versions (King James Bible, American Standard Version) produced practically indistinguishable results.

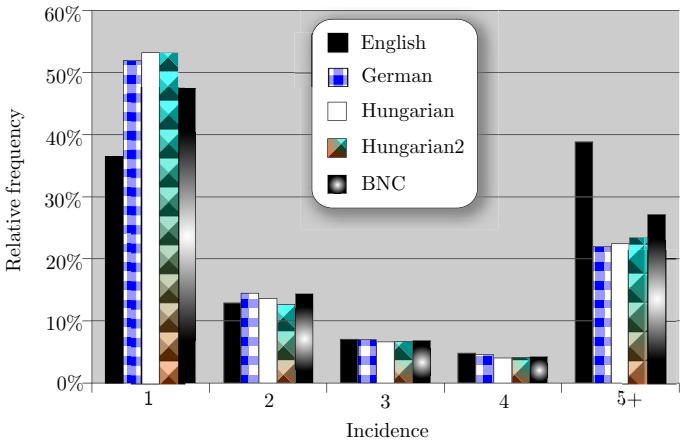

Fig. 3
Frequency of occurrences

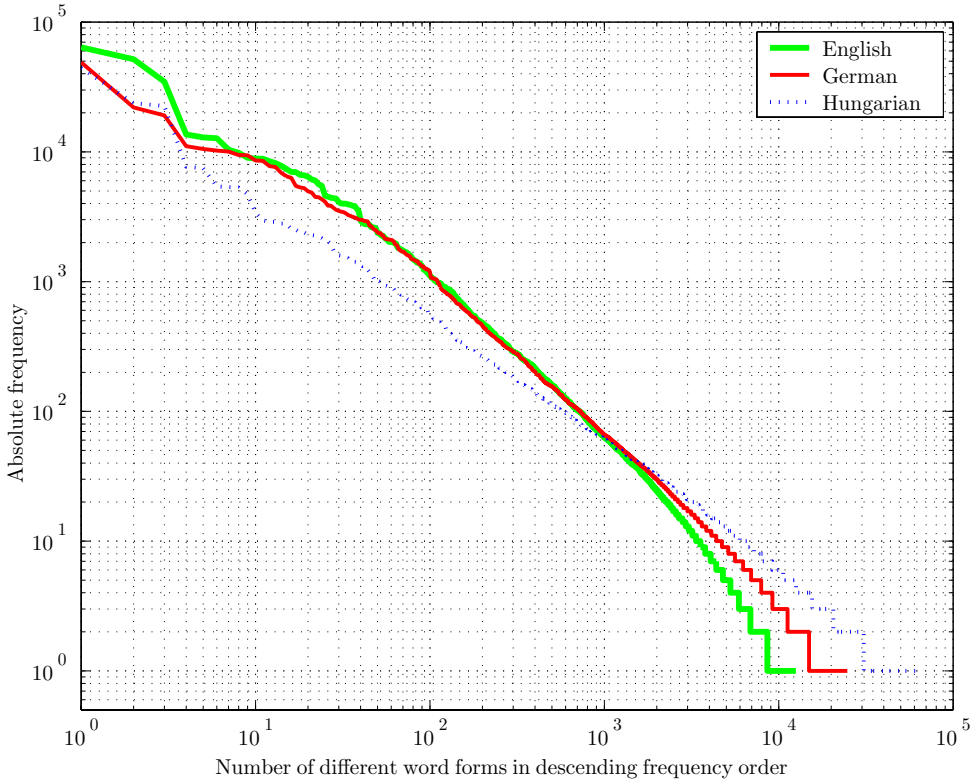

Fig. 4
Ranking of Bibles

The coverage-rank distribution of the Bible versions shows very similar properties to the general texts in English and German. The small size of the corpus results in a very distorted function shape for Hungarian.

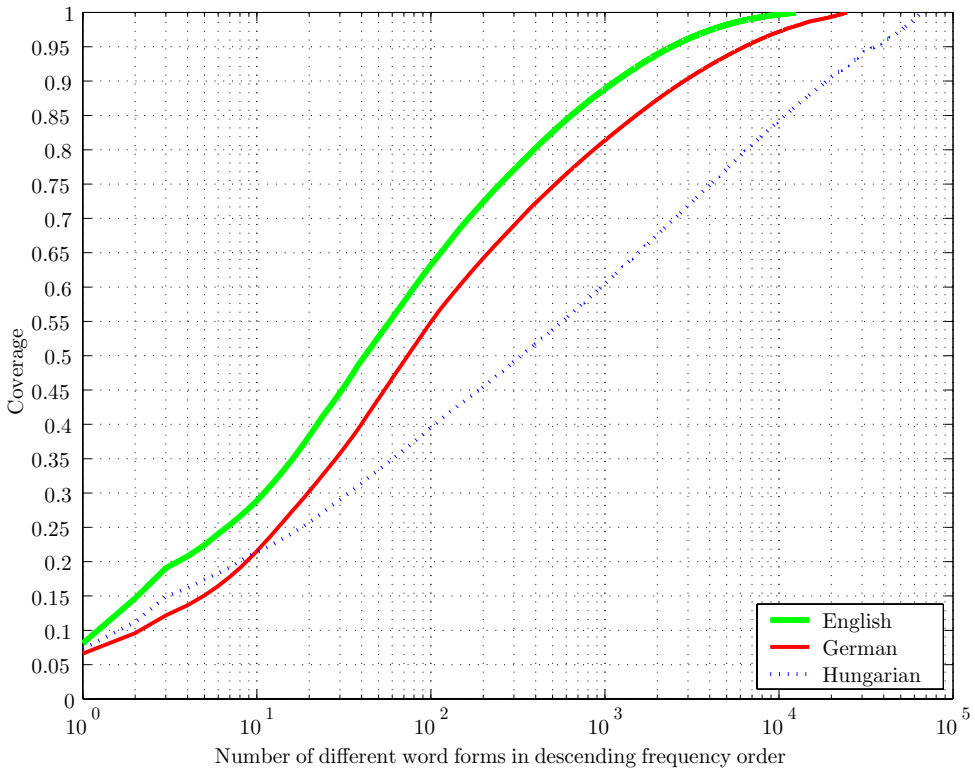

Fig. 5
Coverage of Bibles

3.3. Word length

Figure 6 (overleaf) gives the word length distributions of our corpora. Lines labelled by W. are weighted distributions (i.e., every word is counted) while the "normal" distributions are calculated from the list of different words. Average values are given in Table 3 (also overleaf). Although word length is an important factor in several domains, we found only one paper (Sojka 8) jointly dealing with word length distribution of English, German and a highly inflecting language, Czech.

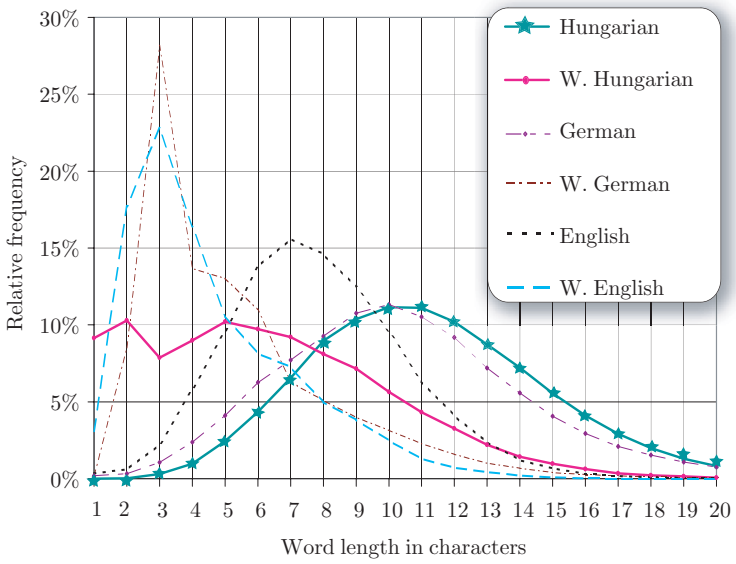

Fig. 6
Statistical distributions of word length

Table 3
Average word length of the corpora (in characters)

LANGUAGE	AVERAGE WORD LENGTH		
	unweighted	weighted	Sojka's results
English	7.85	4.57	8.93
German	10.52	5.29	13.24
Hungarian	11.21	6.24	10.55 (Czech)

The main topic of that paper was compound word hyphenation and word lists were generated from stems by rules. The size of corpora was quite small for English (123,000) and German (368,000) and greater for Czech (3.3M). It is interesting that both the distributions and the average values are very near to ours that come from real running text. The similar results for Czech (Slavic) and Hungarian (Finno-Ugric) are surprising because—besides both being an inflecting language—they have very little in common.

In most practical applications weighted distributions are of greater importance, which **greatly differ** from the "normal" ones (e.g., the "normal" German distribution is nearly identical to Hungarian while the weighted one approximates English).

3.4. Sentence statistics

In this section variability of text at easy-to-detect prosodic boundaries (sentence beginning and end, preceding and following commas) is described according to sentence types (statement, question and exclamation) for the three languages studied. Commas and sentence final characters (., !, ?) were used as signs for prosodic boundaries. In our approach listings, for example, are regarded as separate prosodic units. Special word-like units (e.g., abbreviations, numbers, Roman numbers, etc.) were excluded from the occurrence calculations because they could have distorted the results.

It is clear that this approach does not yield perfect results in the narrow grammatical sense. That would require at least a syntactic analyser in a unified framework for the three languages studied. Such a tool is not available for us. It is not important in our case to find all boundaries, we rather concentrate on finding several boundaries which are expressed in human reading. We suppose that our labelling provides such unit boundaries. That was confirmed by visual inspection of corpus samples in the three languages. All corpora contain about an equal number of sentences in each sentence type per language, statements being approximately 10 times as frequent as questions and exclamations. The number of statements is between 114,000 and 134,000 while the number of questions and exclamations varies between 9,000 and 16,000. Further numerical results are given in tables of the Appendix. Our textual analysis is based on bar-graphs with the aim of easier comprehension.

Average word frequency is defined as the ratio of the total number of all analysed words in a position and the number of different words found in the same position. It gives the average value of a word being re-used in a given position. Figure 7 (overleaf) describes the results for statements, questions and exclamations. Five word unit categories (first and last in a sentence, preceding and following commas and the remaining positions) are analysed for the three sentence types.

It seems that English and German statements have a very similar distribution. A word is used nearly twenty times on the average in the initial positions of prosodic units (first in sentence and following comma). The outstandingly high value for the 'other position' column of English statements might be the result of the significantly smaller vocabulary size (see Figure 2). Re-usage of words at the final position of prosodic units (last in sentence and preceding comma) is somewhat higher for English than for German. It is interesting to note that only English and German statements display higher regularity preceding a comma than at the end of the sentence.

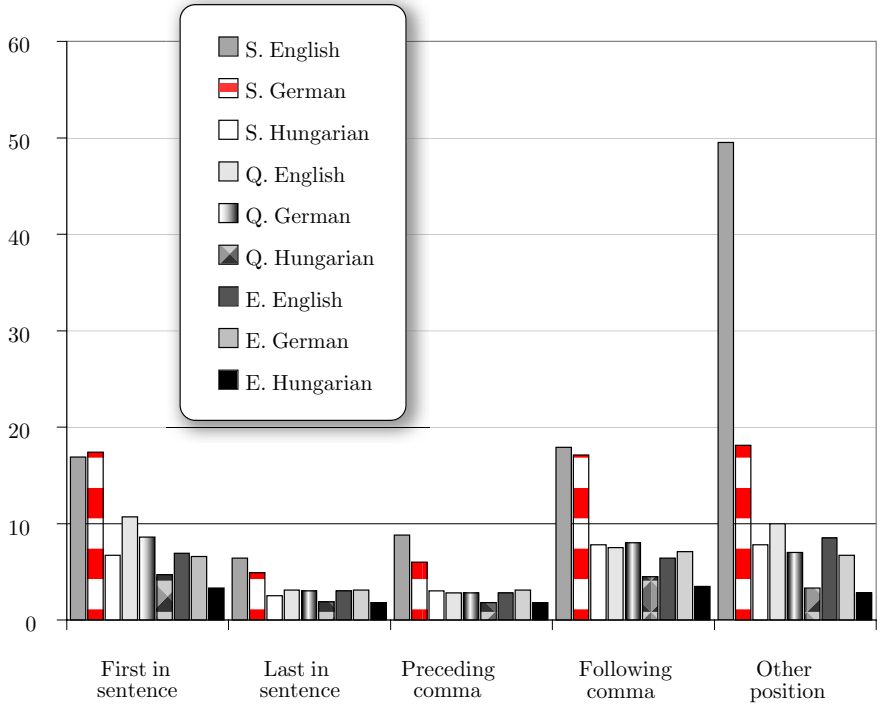

Fig. 7
Average word frequency of statements (S), questions (Q) and exclamations (E) in the positions studied

English and German questions also display some similarities. It is surprising that re-usage values are much smaller for questions than for statements. The general opinion is that questions are more structured than statements. Our results contradict this assumption. The difference is the smallest in Hungarian. It is also interesting that German questions after commas have a higher re-usage value that in English. Re-usage at the end of units is very small for all the three languages.

Exclamations are characterised by the smallest amount of re-usage. Values for English and German are very similar. Re-usage values in 'first in sentence' and 'following comma' positions are nearly the same. Hungarian has smaller values but features of the above-mentioned positions are nearly identical.

The values for Hungarian word re-usage are approximately 50% of those for the other languages. This is due to the much higher number of word forms (see e.g., Tables 1 and 2, Figure 2) which is the consequence of the agglutinative nature of the language. This result is in line with our preliminary assumptions.

4. Applications

4.1. Language detection

The language detection module was developed for an experimental e-mail reading system. A study of real-life e-mail data showed that approximately 40% of the e-mail which is passing through a Hungarian Internet provider contain English only messages. About 56% contains mainly Hungarian text. German e-mail accounts for approximately 3%. This situation requires accurate, automatic prediction of the language(s) used in a given e-mail. The first step for language determination was the creation of word frequency lists from our text corpora. As an example of the results, the twenty most frequent words are given in Table 4 (overleaf). The 100% value is associated with the most frequently used word in a language. The frequency of other words is shown in proportional relationship to this base value.

The data, provided by word frequency lists, cannot be used directly for the detection of languages because there are character combinations appearing in more than one language. Such examples are shown by bold characters in Table 4, e.g., *is* and *in*. Starting from word frequency lists, in the next step a final **list of keywords** was created for each language and these lists were used for language detection. The following rules have been set up for the detection of the language (English, German and Hungarian) of a sentence:

(4) (a) Any sentence containing at least one Hungarian keyword is Hungarian.

(b) The sentence is English if it contains no Hungarian keywords and there are fewer German keywords than English.

(c) The sentence is German if it contains no Hungarian keywords and there are fewer English keywords than German.

(d) If there were no language-related keywords in the sentence, then the language of the sentence is the same as the one which preceded it.

(e) If there were no language-related keywords in the sentence and there is no previous sentence, the sentence is by default Hungarian.

Table 4

The first 20 words of the relative frequency list in three languages

	HUNGARIAN		ENGLISH		GERMAN	
	Word	Freq. (%)	Word	Freq. (%)	Word	Freq. (%)
1	**a**	100.0	the	100.0	und	100.0
2	az	35.54	of	49.50	die	72.39
3	hogy	18.20	and	35.15	der	69.97
4	s	16.42	to	29.81	sie	49.16
5	nem	15.76	**a**	26.21	das	43.09
6	**is**	12.05	his	21.58	er	42.79
7	és	10.82	**in**	20.35	es	36.75
8	egy	9.872	he	14.16	**in**	31.93
9	volt	7.457	that	12.82	war	31.51
10	meg	6.845	was	11.29	den	28.86
11	azt	6.307	with	10.31	ich	24.18
12	csak	5.716	as	9.25	ein	23.92
13	de	5.713	their	8.64	nicht	22.51
14	ez	4.837	it	8.58	zu	22.01
15	van	4.666	had	8.41	aber	20.76
16	ha	4.470	by	7.51	dem	19.63
17	már	4.367	on	6.66	auf	19.42
18	még	4.044	**is**	6.54	mit	18.52
19	el	3.553	which	6.49	so	18.06
20	mint	3.525	for	6.36	sich	16.66

It can be seen from the rules in (4) that Hungarian has a preference over foreign languages in this system. For this reason, the Hungarian keywords had to be selected carefully. The number of keywords for English and German had to be approximately the same, to ensure equal detection probability. Let us see an example for possible wrong language detection: My aunt said "Mein Freund hat im Januar Geburtstag". If only the underlined English words are included in the vocabulary while the German words are not there, the sentence may be incorrectly labelled as English. Such mixed-language sentences need further processing to be developed later.

The current keyword list contains only 97 Hungarian items (because of rules 1 and 5) together with 172 English and 162 German items. The Hungarian section contains two forms, with diacritics and without, as language detection precedes diacritic placement. The accuracy of correct detection of sentences containing more than 10 characters is approximately 96%.

4.2. Diacritic regeneration

Letters with diacritics are represented in Hungarian e-mails in several forms but in most cases incorrectly.

Table 5

Possible forms of vowels with and without diacritics

VOWELS WITHOUT DIACRITICS	CORRECT FORMS OF THE 14 VOWELS	NO. OF VERSIONS
a	a á	2
e	e é	2
i	i í	2
o	o ó ö ő	4
u	u ú ü ű	4

Table 5 gives the possible combinations of the *a, e, i, o, u* characters in Hungarian e-mails. With increasing number of vowels in a word, the number of possible variations increases quickly.

Table 6

Possible forms with different diacritic positions for the word *veres* 'red'

veres	**Veres** az ég. (The sky is **red**)
verés, véres	A **verés** után ***véres*** lett a háta. (After the **beating** his back became ***bloody***). or A **verés** után ***veres*** lett a háta. (After the **beating** his back became ***red***).

Table 6 illustrates the three possible meanings which could be generated from the ASCII character string *veres*. The number of possible forms for giving diacritics in a word is given by the formula in (5) where *vno* is the number of vowels in the word:

(5) $2^{vno} \leq word\_forms \leq 4^{vno}$

An example is given in (6); the correct form with diacritics is *megbízhatósága* 'its reliability'.

(6)

m	e	g	b	i	z	h	a	t	o	s	a	g	a
	2			2			2		4		2		2

So this word may be provided with diacritics in $2 \times 2 \times 2 \times 4 \times 2 \times 2$ theoretical forms and only one is correct.

For Hungarian there was no diacritic retrieval software available before the development of our experimental e-mail reading system. Therefore, efforts have been made to combine different available software with our new algorithms for diacritic retrieval. In the experimental e-mail reader version, several combinations have been tested. In the industrial system a version that is based on vocabularies derived from statistically processed Hungarian text corpora was used. The diacritic regeneration vocabulary was derived by processing approximately 80 million words. Words with diacritics (+D) were indexed by their pairs without diacritics (−D). If more than one word had the same −D form, the most frequent +D version was included in the vocabulary. The final vocabulary contains approximately 1.5 million different word forms. It is important to note that this rather large number is the result of the agglutinative nature of Hungarian. Even with this large vocabulary, the probability of correct diacritic regeneration is only around 96%.

In order to reach a real-time solution, in the industrial version only this vocabulary-based solution could be used because more complex linguistic analysis modules were too slow. The memory requirement of the vocabulary database is approximately 40 Mbytes.

4.3. Implications for Hungarian speech generation

Speech generation using unlimited vocabularies is not a traditional text-to-speech problem anymore. The language of the text may be unknown, several types of errors (typing errors, lack of diacritics, etc.) might appear in several applications (e-mail reading, news, books, etc.). Users are far more sensitive to errors in the auditory channel than in the visual way. In the latter case during education automatic correction mechanisms are built out.

It is of utmost importance for unlimited vocabulary synthesis that the input text should be as well defined and correct as possible. In order to reach this aim, the error rate of automatic correction mechanisms should be further reduced. New message types (e.g., SMS) present further problems for reading. In these cases users frequently create special abbreviations which are often not known to the general public, so there is no way to prepare reading systems for them. Statistical and rule-based methods might be used for detecting these special text types and present them for interpretation to a human operator.

Our results also show that fully word unit based concatenative speech synthesis is not practical even for English. It was also shown, however, that

in all languages prosodic boundaries are marked by frequently used keywords (cf. Figure 7 and tables in the Appendix). It seems to be reasonable to predict that using naturally pronounced units in these positions might improve the overall impression of the listener.

5. Conclusions

- All corpora show a very similar coverage distribution which can be well approximated by straight lines on a logarithmic scale (i.e., the number of different word forms exponentially grows if higher text coverage is to be achieved). The original Zipf's law and its improvements are sufficiently correct in a very limited range only.

- The Hungarian vocabulary size is about 5 times greater than German and 20 times greater than English in a corpus of similar **coverage** distribution. If the **size** of the Hungarian corpus is similar to the others (i.e., coverage is smaller) this decreases to 2 and 5, respectively.

- A single measure for comparing language complexity in case of corpora of similar size is proposed. We propose the number of words needed for covering 95% of a sufficiently large, representative corpus of a language. The corpus should be regarded as representative, if it reaches the saturation state above 95%. The name of the measure could be COV95(corpus_size,language). Study of further languages is proposed to verify the usefulness of the proposed measure.

- For Hungarian and German more than 50% of corpus elements appeared only once, which make advance closed training of real-life large vocabulary applications practically impossible.

- "Normal" and weighted word length distributions greatly differ, the average is approximately halved.

- All languages exhibit similarities in the relative structural importance of the five prosodic boundary positions. German and English re-uses word forms to a similar extent in these positions which is about the double of the values for Hungarian.

- Practical open vocabulary applications need to incorporate rule-based linguistic knowledge if the application is complex and/or the error rate

should be low. Language detection and diacritic regeneration applications of our statistical text analysis have been described with an error rate of 5–6%. Smaller error rates can be achieved by increasing the corpus size and including rule based corrections for word sense disambiguation and similar problems.

The results can be applied in such diverse domains as predictive text input, diacritic regeneration from 7bit ASCII unaccented forms, word hyphenation, language modelling in speech recognition, corpus-based speech synthesis, etc. Related aspects of an e-mail reading application are described in detail in Németh et al. (5).

Appendix

Table A1

Hungarian sentence statistics

	HUNGARIAN	NO. OF WORDS ANALYSED	DIFFERENT WORDS	AVERAGE WORD FREQ.	THIS CAT. / ALL DIFF. WORDS
STATEMENT	First in sentence	132411	19843	6.7	7.5%
	Last in sentence	132123	52358	2.5	19.8%
	Preceding comma	253887	84001	3.0	31.8%
	Following comma	231739	29862	7.8	11.3%
	Other position	1555475	198742	7.8	75.2%
	Distribution ratio		1.46		
	Full sub-corpus	2305635	264415	8.7	100.0%
QUESTION	First in sentence	12446	2661	4.7	8.4%
	Last in sentence	12441	6541	1.9	20.6%
	Preceding comma	13520	7408	1.8	23.3%
	Following comma	11632	2612	4.5	8.2%
	Other position	71050	21831	3.3	68.8%
	Distribution ratio		1.29		
	Full sub-corpus	121089	31729	3.8	100.0%
EXCLAMATION	First in sentence	11192	3370	3.3	11.3%
	Last in sentence	11175	6120	1.8	20.5%
	Preceding comma	14117	7905	1.8	26.4%
	Following comma	11423	3264	3.5	10.9%
	Other position	54246	19053	2.8	63.7%
	Distribution ratio		1.33		
	Full sub-corpus	102153	29909	3.4	100.0%
ALTOGETHER		2516648	281214	8.9	

Each table contains data for a particular language. Five word unit categories (first and last in a sentence, preceding and following commas and the remain-

ing positions) are analysed for the three sentence types. The sentence type is given in the 1st column. The first 5 rows for each sentence type give statistics for the given position. Row 6 for a sentence type is the ratio of the sum of the different words in the 5 positions and the number of different words in the given sub-corpus (e.g., 6764+17926+29692+13145+49872/60469 yield 1.94 in Table A3). Row 7 contains information related to the full sub-corpus of the given sentence type. The last row of each table contains total values for the given language. Column 2 contains short reminders to data types. Column 3 gives the total number of analysed words found in a certain position (it is equal to the number of sentences of the given sentence type). Column 4 contains the number of different words in a position of a sentence type. Column 5 gives the average number of use of a word in a given position (ratio of column 3 and 4). Column 6 is the ratio of column 4 and the number of different words in a sentence type (column 4, row 7). The percentage values of column 6 of a sentence type do not sum up to 100% because the same word of the corpus might appear in several positions.

Table A2

German sentence statistics

	GERMAN	NO. OF WORDS ANALYSED	DIFFERENT WORDS	AVERAGE WORD FREQ.	THIS CAT. / ALL DIFF. WORDS
STATEMENT	First in sentence	133462	7659	17.4	5.6%
	Last in sentence	133420	26970	4.9	19.8%
	Preceding comma	258889	43485	6.0	32.0%
	Following comma	247174	14497	17.1	10.7%
	Other position	2002630	110602	18.1	81.4%
	Distribution ratio		1.50		
	Full sub-corpus	2775575	135924	20.4	100.0%
QUESTION	First in sentence	13976	1625	8.6	7.6%
	Last in sentence	13975	4637	3.0	21.8%
	Preceding comma	16599	5884	2.8	27.6%
	Following comma	14793	1838	8.0	8.6%
	Other position	112794	16134	7.0	75.8%
	Distribution ratio		1.41		
	Full sub-corpus	172137	21291	8.1	100.0%
EXCLAMATION	First in sentence	16029	2420	6.6	10.8%
	Last in sentence	16012	5243	3.1	23.4%
	Preceding comma	19779	6474	3.1	28.9%
	Following comma	16636	2344	7.1	10.4%
	Other position	111618	16552	6.7	73.8%
	Distribution ratio		1.47		
	Full sub-corpus	180074	22440	8.0	100.0%
ALTOGETHER		3117661	143778	21.7	

Table A3

English sentence statistics

	ENGLISH	NO. OF WORDS ANALYSED	DIFFERENT WORDS	AVERAGE WORD FREQ.	THIS CAT. / ALL DIFF. WORDS
STATEMENT	First in sentence	114410	6764	16.9	11.2%
	Last in sentence	114292	17926	6.4	29.6%
	Preceding comma	261544	29692	8.8	49.1%
	Following comma	235750	13145	17.9	21.7%
	Other position	2470432	49872	49.5	82.5%
	Distribution ratio		1.94		0.0%
	Full sub-corpus	3196428	60469	52.9	100.0%
QUESTION	First in sentence	9228	863	10.7	6.7%
	Last in sentence	9228	3019	3.1	23.4%
	Preceding comma	11304	4018	2.8	31.1%
	Following comma	9742	1299	7.5	10.1%
	Other position	102012	10232	10.0	79.2%
	Distribution ratio		1.50		
	Full sub-corpus	141514	12912	11.0	100.0%
EXCLAMATION	First in sentence	8816	1282	6.9	9.8%
	Last in sentence	8812	2940	3.0	22.4%
	Preceding comma	12101	4309	2.8	32.8%
	Following comma	10169	1600	6.4	12.2%
	Other position	86513	10137	8.5	77.2%
	Distribution ratio		1.54		
	Full sub-corpus	126411	13131	9.6	100.0%
ALTOGETHER		3458856	62501	55.3	

References

[am] American Standard Version of the Bible.
(HTTP://EBIBLE.ORG/BIBLE/ASV)

[dig] Digital Library Academy.
(HTTP://ALFRED.NEUMANN-HAZ.HU)

[elb] Elberfelder Bible.
(HTTP://HEILIGE-SCHRIFT.SYTES.NET)

[1] Gibbon, Dafydd – Roger Moore – Richard Winski 1998. Spoken language characterisation. Mouton de Gruyter, The Hague.

[gutp] Gutenberg Project.
(HTTP://WWW.GUTENBERG.AOL.DE)

[hel] Hungarian Electronic Library.
(HTTP://WWW.MEK.IIF.HU)

[kat] Katolikus Biblia.
(HTTP://WWW.EXTRA.HU/SZENTIRAS)

[2] Kilgarriff, Adam 2002. BNC database and word frequency lists.
(HTTP://WWW.ITRI.BTON.AC.UK/~ADAM.KILGARRIFF/BNC-README.HTML)

[king] King James Bible.
(HTTP://WWW2.CCIM.ORG/BIBLE/DCB.HTML)

[3] Li, Wentian 2002. Bibliography of references to Zipf's law.
(HTTP://LINKAGE.ROCKEFELLER.EDU/WLI/ZIPF/)

[4] Németh, Géza – Csaba Zainkó 2001. Word unit based multilingual comparative analysis of text corpora. In: Proceedings of Eurospeech 2001, 2035–8. Aalborg, Denmark.

[5] Németh, Géza – Csaba Zainkó – László Fekete – Gábor Olaszy – Gábor Endrédi – Péter Olaszi – Géza Kiss – Péter Kis 2000. The design, implementation, and operation of a Hungarian e-mail reader. In: International Journal of Speech Technology 3 : 217–36.

[6] Popescu, Ioan-Iovitz 2002. On the Lavalette's nonlinear Zipf's law.
(HTTP://WWW.GEOCITIES.COM/IIPOPESCU/ZIPFS_LAW.HTML)

[7] Roukos, Salim 1996. Language representation. In: Ronald A. Cole – Joseph Mariani – Hans Uszkoreit – Annie Zaenen – Victor Zue (eds) Survey of state of the art in human language technologies. Cambridge Univeristy Press, Cambridge.
(HTTP://CSLU.CSE.OGI.EDU/HLTSURVEY/CH1NODE8.HTML#SECTION16)

[8] Sojka, Petr 1995. Notes on compound word hyphenation in TEX. In: Proceedings of TUG'95, September 1995, 290–6.

[9] Váradi, Tamás 1999. On developing the Hungarian National Corpus. In: Špela Vintar (ed.) Proceedings of the Workshop Language Technologies—Multilingual Aspects, 32nd Annual Meeting of the Societas Linguistica Europea, Ljubjana, Slovenia, 57–63. Faculty of Arts, University of Ljubljana, Ljubjana.

Address of the authors: Géza Németh – Csaba Zainkó
Department of Telecommunications and Telematics
Budapest University of Technology and Economics
Magyar tudósok körútja 2.
H–1117 Budapest
{nemeth, zainko}@ttt.bme.hu