Word Unit Based Multilingual Comparative Analysis of Text Corpora

Géza Németh and Csaba Zainkó

Department of Telecommunications & Telematics
Budapest University of Technology and Economics, Hungary
{nemeth, zainko}@ttt.bme.hu

Abstract

Parallel study of three very 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 was used. The corpus size was the same (app. 20Mbytes, 2.5-3.5 million word forms) for all languages. Besides traditional corpus coverage, word length and occurence statistics, some new features about prosodic boundaries (sentence beginning 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. The functions are much nearer for English and German than for Hungarian. The results can be applied in such diverse domains as predictive text input, word hyphenation, language modeling in speech recognition, corpus-based speech synthesis, etc.

Keywords: text corpora, corpus analysis, multilinguality, word length, unit based analysis, language modeling, corpus-based speech synthesis

1. Introduction

As language and speech technology applications gain an increasingly wide-spread use in several languages/countries, it is important to re-examine whether how much difference exists between English (in most cases the first language for most 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.

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. Besides traditional corpus coverage (Figure 3 -in order to improve readability it was placed at the end of the paper-, Table 1, 2), occurence statistics (Figure 1) and weighted and unweighted word length (Figure 2, Table 3), some new features about prosodic boundaries (sentence beginning and final positions, preceding and following a comma, Table 5-7) were also computed.

2. Text corpora

Corpora of publicly available Internet sources was used. The corpus size was the same (app. 20Mbytes, 2.5-3.5 million word forms) for all languages. 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 not 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 held are numbers, Roman numbers and abbreviations.

The Hungarian corpus was selected from texts larger than 50kbytes in the Hungarian Electronic Library [1] (HEL, app. 2.5 million words). The German corpus was collected from similar material of the Gutenberg project [2] (app. 3.1 million words). The English corpus was collected from English sections of HEL (app. 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. In order to compare coverage effects, a larger corpus (denoted by Hungarian2) was generated for Hungarian by adding data to the HEL corpus from online newspapers and combining it with a list of 700 kwords which was derived from up-to-date texts containing 21 million words (Hungarian National Corpus [3]). Hungarian2 contains app. 1.1 million different words.

3. Discussion

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. [4] and [5]). 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 required most frequent words by static coverage

Language	Static coverage		
	75%	90%	97,5%
Hungarian	10650	70000	400000
German	2000	14550	80000
English	1250	5800	20100

Table 2. gives the coverage rate using the 1000, 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 static coverage

Language	Number of most frequent words		
	1000	20000	100000
Hungarian	51.8%	80.7%	92.0%
German	69.1%	91.8%	98.1%
English	72.8%	97.5%	(100%)

The coverage percentage as a function of the most frequent words set in a descending order. are given in Figure 3. It can be found at the end of the paper because of its' large size for ease of readability. The vertical axis is linear while the horizontal one is logarithmic in order to ensure a display ratio of 1 - 1.000.000. It is an interesting result, that in the 40 - 85% range all functions run nearly parallel and could be well approximated by straight lines (i.e. *the relationship is nearly purely exponential*).

It seems, that 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 awaited according to theoretical assumptions. 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 app. doubling the vocabulary -36k to 75k- one can jump from 85% to 90%, but increasing it from 166k to 374k raises coverage from 95% to 97% only). Even this section could be well approximated by straight lines on the figure.

It is also worth looking at the lower end of the figure. The 10 most frequent words cover 20 - 30%, the first 100 cover 30 - 50% while the first 1k provide 50 - 70% coverage. It 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 1. illustrates a very problematic aspect of corpus based approaches. It is clear, that even for English, which contained only 62k different word forms in a 3.5 million corpus, nearly 40% of the 62k 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*. 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.

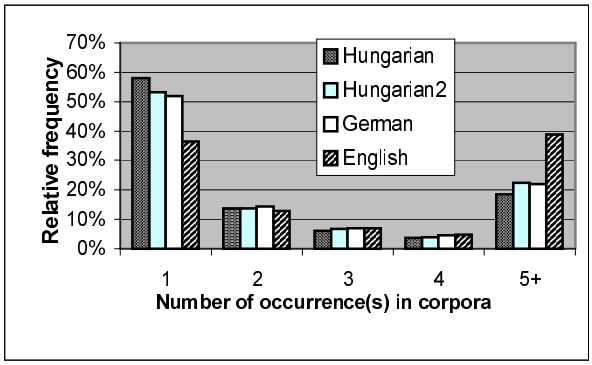

Figure 1: Frequency of occurrences

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

3.2. Word length

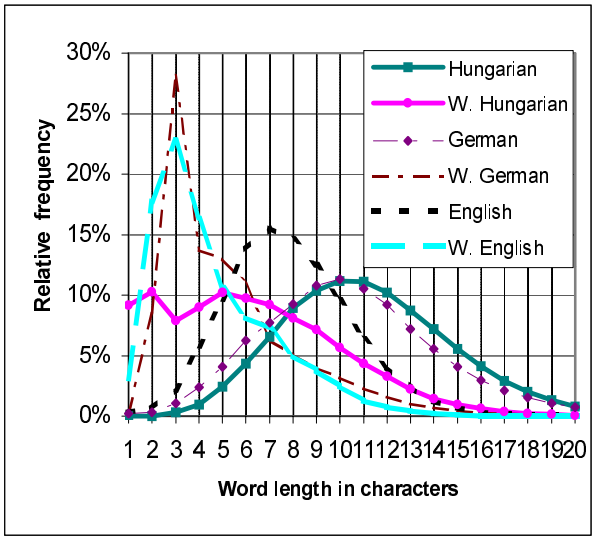

Figure 2: Statistical distributions of word length

Figure 2. gives the word length distributions of our corpora. Lines denoted 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. Although word length is an important factor in several domains, we found only one paper -written by Sojka- jointly dealing with word length distribution of English, German and a highly inflecting languge (Czech, [6]).

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

Language	Average word length		
	unweighted	weighted	Sojka's results
Hungarian	11,21	6,24	10,55 (Czech)
German	10,52	5,29	13,24
English	7,85	4,57	8,93

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 (123k) and German (368k) 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 there is 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).

3.3. 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). Special word-like units (e.g. abbreviations, numbers, Roman numbers, etc.) were excluded from the occurrence calculations that's why numbers that should logically be equal (e.g. no. of first and last words in Hungarian statements, Table 5, first two numbers in the No. of words column) may differ in Table 5-7, that give the results.

Each table contains data in a particular language. Five word unit categories (first and last in a sentence, preceding and following commas and the remaining positions) are analyzed 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 7). 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 no. of analyzed 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 no. 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

Table 5: Hungarian sentence statistics

	Hungarian	No. of words analyzed	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	

Table 6: German sentence statistics

	German	No. of words analyzed	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 7: English sentence statistics

	English	No. of words analyzed	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	

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.

All corpora contain about equal number of sentences in sentence types, statements being app. 10 times as frequent as questions and exclamations. English and German show similar values. It is interesting -however- that statements much more frequently use the same starting word (c.f. column 5) than the other sentence types. The values for Hungarian word re-usage (column 5) are app. 50% of those for the other languages.

4. 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 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
- For Hungarian and German more than 50% of corpus elements appeared only once, which make advance closed traning of real-life large vocabulary applications practically impossible
- "Normal" and weighted word length distributions greatly differ, the average is app. halved
- All languages exhibit similarities in the relative structural importance of the 5 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

The results can be applied in such diverse domains as predictive text input, diacritic re-generation from 7bit ASCII unaccented forms, word hyphenation, language modeling in speech recognition, corpus-based speech synthesis, etc. Related aspects of an e-mail reading application are described in [7].

5. Acknowledgements

The authors are thankful for the help of Manuel Kaesz in collecting the text corpora of equal size.

6. References

[1] Hungarian Electronic Library, *http://www.mek.iif.hu*
[2] Gutenberg Project *http://gutenberg.aol.de*
[3] Váradi, T., "On Developing the Hungarian National Corpus", in Vintar, S. (ed.): *Proceedings of the Workshop Language Technologies - Multilingual Aspects*, 32nd Annual Meeting of the Societas Linguistica Europea, Ljubjana, Slovenia, pp. 57 - 63
[4] Gibbon, D., Moore, R., and Winski, R., *Spoken Laguage Characterisation* Mouton de Gruyter, 1998, pp. 41-45
[5] Salim Roukos, "Language Representation", in Cole et. al., *Survey of State of the Art in Human Language Technologies* (1996), http://cslu.cse.ogi.edu/HLTsurvey/ch1node8.html#SECTION16
[6] Petr Sojka, ”Notes on compound word hyphenation in TeX”, *Proc. of TUG'95*, September 1995, pp. 290-296
[7] Németh, G., Zainkó, Cs., Fekete, L., Olaszy, G., Endrédi, G., Olaszi, P., Kiss, G., Kis, P., ”The Design, Implementation, and Operation of a Hungarian E-Mail Reader” *International Journal of Speech Technology*, Kluwer Academic Publishers, Volume 3, Nos 3/4, December 2000, pp. 217-236

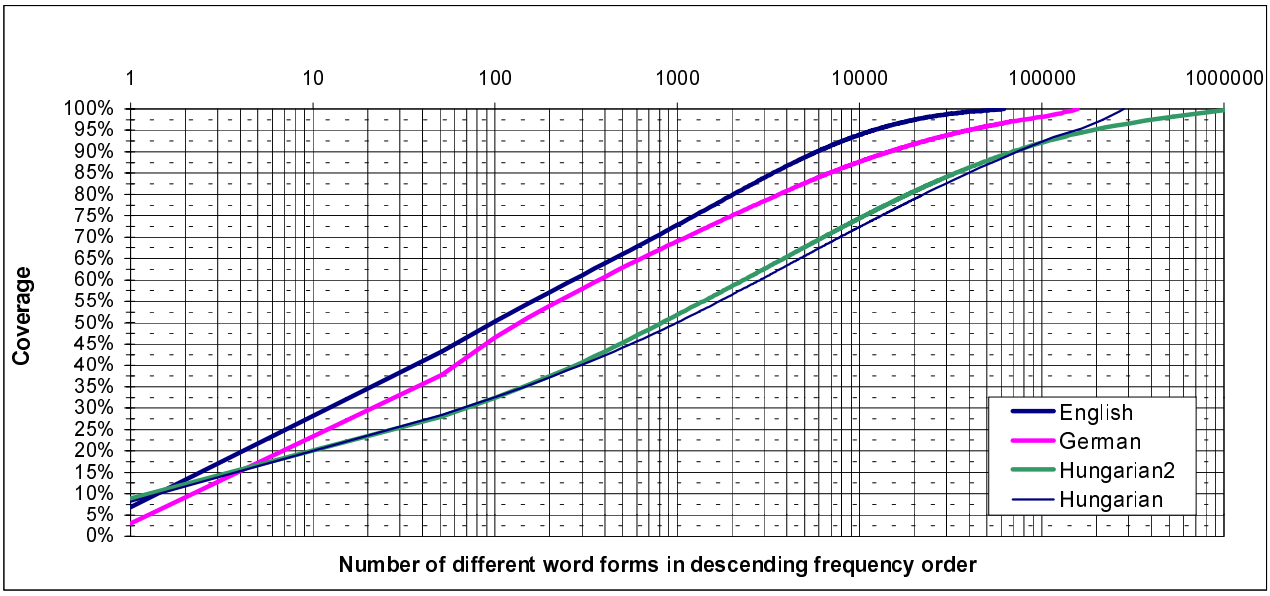

Figure 3: Corpora coverage by the most frequent words (logarithmic horizontal scale)