Automatic Recognition of Czech Derivational Prefixes

Alfonso Medina Urrea[1] and Jaroslava Hlaváčová[2]

[1] GIL IINGEN UNAM,
Apartado Postal 70-472, 04510 Coyoacán, DF, Mexico
amedinau@iingen.unam.mx
[2] ÚFAL MFF UK
Malostranské náměstí 25, 11800, Praha, Czech Republic
hlava@ufal.mff.cuni.cz

Abstract. This paper describes the application of a method for the automatic, unsupervised recognition of derivational prefixes of Czech words. The technique combines two statistical measures — Entropy and the Economy Principle. The data were taken from the list of almost 170 000 lemmas of the Czech National Corpus

1 Introduction

Our contribution concerns only those languages where words are created by means of affixes. Usually, there exists a quite stable vocabulary, but it is possible to create entirely new words adding suffixes and/or prefixes to already existing ones. If the derivation follows common rules for word creation, everybody understands them, even if they have never seen them before.

The Czech language belongs to the group of languages that derive their vocabulary mainly by means of adding affixes. While the set of suffixes is very stable and does not change during long periods of time, prefixes are much more vivid. Of course, there is a set of old, traditional prefixes, that have been used for a very long time and do not change. But one can very easily add a morph, usually borrowed from other languages, in front of an existing word and create an entirely new word. The old prefixes can be found in every grammar, but the new ones cannot.

Everybody who understands the language understands new prefixes. Everybody but computers. And for any analysis of language, it is very important to know them. Without a sufficiently large list of prefixes, we cannot run successfully enough a morphological analysis, which usually stands on the basis of all automatic language processing.

To be specific — the morphological analyzer always encounters unknown words; that is, words for which it does not recognize their basic forms nor their morphological categories. It is possible to design a "guesser" that uses special properties of the language which could help to guess those basic features of the unknown word. In Czech, we usually take suffixes as the basis [1].

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 189–197, 2005.

The list of prefixes can serve as another type of such guesser: if we get a word that is not included in our morphological dictionary, we would try to see if any of the prefixes matches the beginning of the word. If so, it is probable, that the rest of the word, after tearing the prefix off, will be recognized by the morphological analyzer.

2 Word Sample

The empirical word source for this paper is a list of around 170 000 word types. The basis of our experiment was the Czech National Corpus (CNC) with 100 million word forms. As prefixes do not change with word declensions, we worked with basic word forms — lemmas. There are more than 800 000 different lemmas in the CNC, but the great majority of them have very low frequency. We selected only the lemmas with frequency of at least 5, mainly because words with lower frequencies are very often typos or other rubbish. Their total number is 166 733. In order to make the list smaller, we left out those parts of speech that do not have prefixes, namely prepositions and conjunctions. We took into account only lemmas not beginning with a capital letter, because these are almost 100% proper nouns that usually do not have prefixes.

There are some letters that are untypical for Czech — a Czech word cannot begin with y; letters g, f, w and x are very unusual and there is only a limited number of words containing them. In our list, it would be easy to go through them manually and check whether there is a foreign new prefix or not. However, since the method is unsupervised, we decided not to intervene manually into the process and let the method do it automatically.

3 Method

A full description of the method applied to Spanish can be found in Medina [2]. In that paper, several quantitative measures are explored to compare their desirability as methods to discover affixes. The methods were very successful for suffixes, but not so good for prefixes, surely because in Spanish the former constitute a compact, highly organized inflectional and derivational system, whereas prefixes do not. As the method is general and language independent,[1] we tried to use it for Czech prefixes which, as mentioned above, are very productive.

Let us quickly outline the approach applied. It combines two quantitative methods: measurement of entropy — one of the topics of information theory [4] — and the principle of economy of signs [5, 6]. We will examine some of the reasons why these two methods work well together.

[1] It was applied successfully to a small corpus of Chuj, a Mayan language spoken in Chiapas and Guatemala [3] (essentially with respect to entropy measurement); and recently to a small corpus of Raramuri (Tarahumara), a Yuto-Aztecan language spoken in Northern Mexico (both entropy and economy measurements). Because of space constraints, results for those languages are not presented here.

High entropy measurements have been reported repeatedly as successful indicators of borders between bases and affixes [2, 3, 7, 8, 9, 10]. These measurements are relevant because, as it was pointed out as early as the fifties by linguists like Joseph Greenberg[2] [11], shifts of amounts of information (in the technical sense) can be expected to correspond to the amounts of information that a reader or hearer is bound to obtain from a text or spoken discourse. Frequent segments must contain less information than those occurring rarely. Hence, affixes must accompany those segments of a text (or discourse) which contain the highest amounts of information. And this has been in fact observed for a very wide range of affixes [2, 10, 3], including those whose structural evidence —like that behind the economy principle described below— is not fully provided by a corpus, either because the corpus is too small or not representative of the language [3] or because the affixes in question are old and unproductive [10].

The other important measure used in this experiment is based on the principle of economy of signs (some experiments using measurements based on this principle — either maximum or minimum approaches — are [5, 6, 12, 2, 13]). In essence, for this approach this is a quantity representing how much linguistic structure there is in a given expression. If natural languages are systems, they and their components must be economical to some degree. Thus, we can expect certain signs to be more economical than others because they relate to other signs in an economical way. One aspect of sign economy is evident in that a sign at one level of language, say the lexical one, may be composed of more than one sign of the lower level, say the morphological ones. In this manner, a language can refer at the lexical level to a great number of things using considerably fewer signs than it would be necessary if it had exactly one sign for each thing named.

Affixes can combine with bases to produce a number (virtually infinite) of lexical signs. It is clear that affixes do not combine with every base. Certain ones combine with many bases, others with only a few. Nevertheless, it makes sense to expect more economy where more combinatory possibilities exist.

This refers to the syntagmatic dimension. The paradigmatic dimension can also be considered: affixes alternate in a corpus with other affixes to accompany bases. If there is a relatively small set of alternating signs (paradigms) which adhere to a large set of unfrequent signs (to form syntagms) the relations between the former and the latter must be considered even more economical. This is naturally pertinent for both derivation and inflection; that is, this is as true for lemma affixes, as it is for affixes of the inflected words of discourse.

4 Building a Catalog of Czech Prefixes

The program basically takes the words of the word sample and determines the best segmentation for each one (according to the two measurements discussed

[2] Zellig Harris relied on phoneme counts before and after a given word segmentation (according to a corpus), a matter undoubtedly related to entropy measurement. But he did not specifically refer to information theory, like Greenberg did.

above and described bellow). Each best segmentation represents a hypothesis postulating a base and an prefix. Thus, the presumed prefix (and the values associated with it) are fed into a structure called Catalog. The more frequent a presumed prefix is, the more likely it is really a prefix.[3]

4.1 Information Content

Information content of a set of word fragments is typically measured by applying Shannon's method.[4] As mentioned above, high entropy measurements have been reported repeatedly as successful indicators of borders between bases and affixes.

For this experiment in particular, the task was to measure the entropy of the word segments which follow a prefix candidate, according to the word sample: borders between prefixes and stems tend to exhibit peaks of entropy. Thus, looking for peaks of information meant taking each left-hand substring of each word of the sample, determining the probability of everything that follows, and applying Shannon's formula to obtain an entropy measurement for the right hand substrings related to each left-hand substring examined.

4.2 Economy Principle

The economy of segmentations can be measured by comparing the following sets of word beginnings and endings[5] from a word sample. Given a prefix candidate, there are two groups of word segments:

1. *companions* — word endings which follow the given prefix candidate (syntagmatic relation).
2. *alternants* — word beginnings which alternate (occur in complementary distribution) with the prefix candidate.

The following fraction is a simplified example of how these can be compared to capture the essence of the method proposed in [5, 6, 2, 10]:

$$k = \frac{companions}{alternants} \tag{2}$$

More formally, let $B_{i,j}$ be the set of word endings which follow, according to a corpus, the left-hand word segment $a_{i,j}$, which consists of the first jth letters

[3] Other possibilities, like selecting several best segmentations per word or including some threshold criteria to filter forms with low values, are discussed in Medina [2].

[4] Recall the formula

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

where p_i stands for the relative frequency of word fragment i [4]. See Oakes [9] or Manning and Shütze [14] —among many others— for brief descriptions of the method.

[5] It is worth noting that with the term 'ending' we do not mean here the grammatical ending of a word, but just the substring of letters towards its end.

of the ith word of the corpus. Let $B^s_{i,j}$ be the subset of $B_{i,j}$ consisting of the word endings which are suffixes of the language in question. Let $A^p_{i,j}$ be the set of word beginnings which are, also according to the corpus, prefixes of the language and occur in complementary distribution with the word beginning $a_{i,j}$. One way to estimate the economy of a segmentation between a set word beginnings and a set of word endings, in such a way that the word beginnings are prefixes is:

$$k^p_{i,j} = \frac{|B_{i,j}| - |B^s_{i,j}|}{|A^p_{i,j}|} \tag{3}$$

As established in (2), the numerator of (3) can be described as the set of right-hand *companions* of the left-hand word segment $a_{i,j}$ and the denominator the set of left-hand segments or *alternants* of (in paradigmatic relation to) $a_{i,j}$.

In this way, when an initial word fragment is given, a very large number of companions and a relatively small number of alternants yield a high economy value. Meanwhile, a small number of companions and a large one of alternants indicate a low economy measurement. In the latter case, the word fragment in question is not very likely to represent exactly a morpheme (nor, as we will see, a sequence of them).

4.3 Entropy and Economy Combined

Word segmentation methods can be compared in order to determine how successful they are [7, 2]. But they can also be combined to improve their effectiveness. The methods described above complement each other in the estimation of what can be called the affixality of word fragments. In fact, the values obtained for a given word fragment can be averaged or multiplied. For this experiment, they were normalized and averaged. That is, we estimated *prefixality* by means of the arithmetic average of the relative values of entropy and economy: $(\frac{h_i}{\max h} + \frac{k_i}{\max k}) * \frac{1}{2}$, where h_i stands for the entropy value associated to prefix candidate i; k_i represents the economy measurement associated to the same candidate; and $\max h$ returns the maximum quantity of h calculated for all prefixes (same idea for $\max k$).

The important fact is that the highest values (those expected to occur at the borders between prefixes and bases, and between bases and suffixes) are good criteria to include word fragments as items in the Catalog of Prefixes.

5 Results

The results are shown in Table 1 which contains the ninety prefix candidates with highest affixality values. Candidates are presented in the second column. The third column exhibits frequency of all lemmas from the original word sample, where the candidate comes out as the best prefix. The fourth and fifth columns contain the normalized measurements of entropy and economy. Candidates showing values of less than 0.5 (of either measurement) were filtered.

Table 1. Catalog of Czech Prefixes

	prefix	fr	econ	entr	affty
1.	severo∼	75	0.974	0.93	0.952
2.	proti∼	457	0.928	0.968	0.948
3.	jiho∼	76	0.946	0.922	0.934
4.	mezi∼	199	0.923	0.922	0.922
5.	super∼	263	0.857	0.965	0.911
6.	dvoj∼	233	0.863	0.948	0.905
7.	mimo∼	154	0.879	0.93	0.905
8.	troj∼	136	0.858	0.944	0.901
9.	mnoho∼	103	0.913	0.888	0.901
10.	osmi∼	97	0.929	0.872	0.901
11.	spolu∼	267	0.896	0.902	0.899
12.	video∼	138	0.93	0.868	0.899
13.	východo∼	47	0.926	0.871	0.899
14.	devíti∼	59	0.961	0.833	0.897
15.	při∼	1361	0.910	0.882	0.896
16.	více∼	151	0.886	0.899	0.892
17.	radio∼	102	0.862	0.92	0.891
18.	šesti∼	113	0.93	0.844	0.887
19.	nad∼	437	0.774	1	0.887
20.	celo∼	123	0.871	0.902	0.886
21.	šéf∼	45	0.938	0.833	0.885
22.	pěti∼	168	0.886	0.885	0.885
23.	západo∼	44	0.888	0.881	0.884
24.	sedmi∼	82	0.943	0.817	0.880
25.	několika∼	67	0.957	0.803	0.88
26.	pseudo∼	149	0.82	0.939	0.879
27.	třiceti∼	39	0.944	0.811	0.878
28.	velko∼	172	0.917	0.835	0.876
29.	elektro∼	168	0.802	0.947	0.875
30.	od∼	2393	0.814	0.935	0.875
31.	vy∼	4389	0.838	0.91	0.874
32.	dvanácti∼	51	0.983	0.761	0.872
33.	polo∼	448	0.794	0.95	0.872
34.	středo∼	75	0.849	0.892	0.871
35.	dvaceti∼	48	0.942	0.798	0.870
36.	deseti∼	84	0.895	0.845	0.87
37.	patnácti∼	36	0.968	0.768	0.868
38.	před∼	861	0.803	0.933	0.868
39.	vnitro∼	55	0.894	0.842	0.868
40.	jedno∼	280	0.827	0.908	0.868
41.	tří∼	240	0.867	0.869	0.868
42.	pod∼	1236	0.756	0.977	0.866
43.	dvou∼	304	0.838	0.894	0.866
44.	vysoko∼	52	0.802	0.928	0.865
45.	osmnácti∼	14	0.964	0.761	0.863
46.	horno∼	20	0.887	0.836	0.862
47.	za∼	4052	0.805	0.916	0.860
48.	čtrnácti∼	29	0.939	0.775	0.857
49.	čtyřiceti∼	33	0.92	0.791	0.856
50.	rychlo∼	62	0.856	0.836	0.846
51.	jedenácti∼	24	0.935	0.754	0.845
52.	česko∼	38	0.892	0.792	0.842
53.	foto∼	181	0.787	0.893	0.840
54.	vele∼	84	0.813	0.864	0.839
55.	roz∼	2431	0.769	0.901	0.835
56.	bio∼	164	0.782	0.886	0.834
57.	vodo∼	58	0.782	0.878	0.830
58.	znovu∼	129	1	0.654	0.827
59.	žluto∼	17	0.848	0.804	0.826
60.	mikro∼	256	0.740	0.912	0.826
61.	plno∼	39	0.826	0.826	0.826
62.	nízko∼	54	0.757	0.893	0.825
63.	půl∼	127	0.797	0.853	0.825
64.	roze∼	215	0.851	0.796	0.823
65.	arci∼	31	0.919	0.726	0.823
66.	šedesáti∼	22	0.914	0.730	0.822
67.	na∼	3580	0.730	0.912	0.821
68.	ode∼	126	0.853	0.789	0.821
69.	anti∼	398	0.7	0.939	0.819
70.	malo∼	91	0.781	0.858	0.819
71.	čtvrt∼	55	0.760	0.874	0.817
72.	do∼	2374	0.445	0.924	0.815
73.	staro∼	151	0.742	0.888	0.815
74.	ultra∼	53	0.860	0.769	0.814
75.	euro∼	106	0.722	0.903	0.812
76.	mnoha∼	35	0.881	0.743	0.812
77.	čtyřiadvaceti∼	14	0.881	0.743	0.812
78.	samo∼	229	0.758	0.864	0.811
79.	padesáti∼	39	0.884	0.734	0.809
80.	šestnácti∼	22	0.852	0.766	0.809
81.	modro∼	21	0.775	0.839	0.807
82.	vše∼	169	0.736	0.873	0.805
83.	pů∼	204	0.749	0.858	0.803
84.	osma∼	22	0.886	0.719	0.802
85.	pětiset∼	11	0.820	0.784	0.802
86.	třinácti∼	17	0.911	0.691	0.801
87.	psycho∼	105	0.803	0.795	0.799
88.	popo∼	51	0.836	0.757	0.797
89.	tisíci∼	18	0.901	0.684	0.792
90.	červeno∼	30	0.735	0.849	0.792

The last column contains the affixality index, which was calculated as the arithmetic average of the entropy and economy values of the fourth and fifth columns. Finally, the first column shows the rank of the candidates according to this index.

It is interesting that within the first ninety items there is one segment constituted by two prefixes joined together (*popo-* no. 88). As can be expected, there are more catalog items representing sequences of prefixes down the rest of the Catalog (for example, within the first hundred, *zne-* no. 99).

It is worth noting that there are no false prefixes within the first one hundred candidates (which means that the precision is 1.0 for this set).

In order to calculate recall, a set of productive Czech prefixes was compiled. Thus, the 45 most traditional prefixes were considered in order to determine how many important prefixes the method did not miss. We refer to this set as **T**.

Table 2 shows precision and recall for the first five hundreds of candidates. The first column, labelled **N**, refers to the number of the catalog items considered. The second column, **E**, shows the number of erroneous candidates (mistakes) within the first **N** candidates. The third column shows the precision — (**N-E**)/**N**. The fourth column, **NF**, displays the number of Czech traditional prefixes from the set **T** that were not found within the first **N** candidates (omissions). The recall was calculated as (45-NF)/45. Naturally, the precision decreases with the increasing number of candidates, while the recall exhibits the opposite tendency.

Some of the prefixes are not real prefixes. They could be regarded as word stems that combine with other stems to create new words, by means of composition. However, these (pre)stems behave like prefixes — they are common to more words modifying their meaning. This is, among others, the case of "numerical prefixes" — a special inflective form (usually identical with genitive) of numerals added to a word (mainly adjectives) modifying them numerically. In fact, every number can serve as a prefix in that sense, but it is usually used only for short numerals. If we wanted to say for instance *a dragon that has seven heads* we can say *sedmihlavý drak* — something like *seven headed dragon*, but the *seven headed* corresponds to one word unit in Czech containing a prefix *sedmi* meaning *seven*.

One can object that the prefixes are not divided into groups according to the parts of speech they can join. It is true that some prefixes cannot prefix any word base. That remains an interesting task for future work. In this experiment we just wanted to recognize everything that could serve as a prefix, no matter

Table 2. Evaluation measurements

N	E	precision	NF	recall
100	0	1.0000	23	0.4889
200	12	0.9400	12	0.7333
300	72	0.7600	10	0.7778
400	149	0.6275	7	0.8889
500	229	0.5420	5	0.9556

the context of the rest of the word. The sorting into groups should be a part of a further analysis. Other automatic processings would benefit from it, for instance the guesser mentioned in the Introduction.

6 Conclusions

From the results we can see that it is possible to recognize prefixes independently of the language represented by the corpus (provided they constitute an organized subsystem in that language). There was no false prefix among the first hundred of recognized prefixes. As the list becomes longer (and as the measure of affixality becomes lower), there naturally appear more mistakes.

We can examine how long the Catalog must be in order to be relatively sure that we will have recognized most prefixes. From the very essence of statistics, we can never be sure. But according to the number of wrong prefixes occurring among the items with lower affixality, it seems to us, that 500 would be a good compromise. Although there are probably some more prefixes with lower affixality, their number would be small. As it is always necessary to check the prefixes manually before using them in further analyses, the list should not be too long. We have found that the first 500 items include almost all the traditional prefixes and many new ones.

The comparison of this method with other methods (minimal and maximal distance techniques) is certainly an interesting task for future work. Nevertheless, our approach has shown that it is possible to make a list of prefixes using exact methods. If we had wanted to make the list manually, we would have had to engage in tedious work — searching dictionaries, old grammar books, checking large corpora manually. The method described is useful for everybody who is concerned with morphology of an inflectional language. Moreover, it can recognize even the most modern prefixes that have entered the language quite recently.

Acknowledgments

The work reported on this paper has been supported by a grant of the Czech Ministry of Education LN00A063, by Conacyt's Project R37712A and by DGAPA PAPITT's Project IX402204. Also, thanks to the Institute of the Czech National Corpus for allowing us to use the Czech National Corpus.

References

1. Hlaváčová, J.: "Morphological Guesser of Czech Words". In: Proc. TSD 2001, Berlin Heidelberg, Springer Verlag (2001) 70–75
2. Medina Urrea, A.: "Automatic Discovery of Affixes by Means of a Corpus: A Catalog of Spanish Affixes". Journal of Quantitative Linguistics **7** (2000) 97–114

3. Medina Urrea, A., Buenrostro Díaz, E.C.: "Características cuantitativas de la flexión verbal del chuj". Estudios de Lingüística Aplicada **38** (2003) 15–31
4. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana (1949)
5. de Kock, J., Bossaert, W.: Introducción a la lingüística automática en las lenguas románicas. Volume 202 of Estudios y Ensayos. Gredos, Madrid (1974)
6. de Kock, J., Bossaert, W.: The Morpheme. An Experiment in Quantitative and Computational Linguistics. Van Gorcum, Amsterdam, Madrid (1978)
7. Hafer, M.A., Weiss, S.F.: "Word Segmentation by Letter Successor Varieties". Information Storage and Retrieval **10** (1974) 371–385
8. Frakes, W.B.: "Stemming Algorithms". In: Information Retrieval, Data Structures and Algorithms. Prentice Hall, New Jersey (1992) 131–160
9. Oakes, M.P.: Statistics for Corpus Linguistics. Edinburgh University Press, Edinburgh (1998)
10. Medina Urrea, A.: Investigación cuantitativa de afijos y clíticos del español de México. Glutinometría en el Corpus del Español Mexicano Contemporáneo. PhD thesis, El Colegio de México, Mexico (2003)
11. Greenberg, J.H.: Essays in Linguistics. The University of Chicago Press, Chicago (1957)
12. Goldsmith, J.: "Unsupervised Learning of the Morphology of a Natural Language". Computational Linguistics **27** (2001) 153–198
13. Gelbukh, A., Alexandrov, M., Han, S.Y.: "Detecting Inflection Patterns in Natural Language by Minimization of Morphological Model". In: CIARP-2004. Volume 3287 of Lecture Notes in Computer Science. Springer (2004) 432–438
14. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge (Mass.) (1999)