Web-based frequency dictionaries for medium density languages

András Kornai
MetaCarta Inc.
350 Massachusetts Avenue
Cambridge MA 02139
andras@kornai.com

Péter Halácsy
Media Research and Education Center
Stoczek u. 2
H-1111 Budapest
halacsy@mokk.bme.hu

Viktor Nagy
Institute of Linguistics
Benczúr u 33
H-1399 Budapest
nagyv@nytud.hu

Csaba Oravecz
Institute of Linguistics
Benczúr u 33
H-1399 Budapest
oravecz@nytud.hu

Viktor Trón
U of Edinburgh
2 Buccleuch Place
EH8 9LW Edinburgh
v.tron@ed.ac.uk

Dániel Varga
Media Research and Education Center
Stoczek u. 2
H-1111 Budapest
daniel@mokk.bme.hu

Abstract

Frequency dictionaries play an important role both in psycholinguistic experiment design and in language technology. The paper describes a new, freely available, web-based frequency dictionary of Hungarian that is being used for both purposes, and the language-independent techniques used for creating it.

0 Introduction

In theoretical linguistics introspective grammaticality judgments are often seen as having methodological primacy over conclusions based on what is empirically found in corpora. No doubt the main reason for this is that linguistics often studies phenomena that are not well exemplified in data. For example, in the entire corpus of written English there seems to be only one attested example, not coming from semantics papers, of Bach-Peters sentences, yet the grammaticality (and the preferred reading) of these constructions seems beyond reproach. But from the point of view of the theoretician who claims that quantifier meanings can be computed by repeat substitution, even this one example is one too many, since no such theory can account for the clearly relevant (though barely attested) facts.

In this paper we argue that ordinary corpus size has grown to the point that in some areas of theoretical linguistics, in particular for issues of inflectional morphology, the dichotomy between introspective judgments and empirical observations need no longer be maintained: in this area at least, it is now nearly possible to make the leap from zero observed frequency to zero theoretical probability i.e. ungrammaticality. In many other areas, most notably syntax, this is still untrue, and here we argue that facts of derivational morphology are not yet entirely within the reach of empirical methods. Both for inflectional and derivational morphology we base our conclusions on recent work with a gigaword web-based corpus of Hungarian (Halácsy et al 2004) which goes some way towards fulfilling the goals of the WaCky project (http://wacky.sslmit.unibo.it, see also Lüdeling et al 2005) inasmuch as the infrastructure used in creating it is applicable to other medium-density languages as well. Section 1 describes the creation of the WFDH Web-based Frequency Dictionary of Hungarian from the raw corpus. The critical disambiguation step required for lemmatization is discussed in Section 2, and the theoretical implications are presented in Section 3. The rest of this Introduction is devoted to some terminological clarification and the presentation of the elementary probabilistic model used for psycholinguistic experiment design.

0.1 The range of data

Here we will distinguish three kinds of corpora: *small-, medium-*, and *large-range,* based on the internal coherence of the component parts. A *small-range* corpus is one that is stylistically homogeneous, generally the work of a single author. The largest corpora that we could consider small-range are thus the oeuvres of the most prolific writers, rarely above 1m, and never above 10m words. A *medium-range* corpus is one that remains within the confines of a few text types, even if the authorship of individual documents can be discerned e.g. by detailed study of word usage. The LDC gigaword corpora, composed almost entirely of news (journalistic prose), are from this perspec-

tive medium range. Finally, a *large-range* corpus is one that displays a variety of text types, genres, and styles that approximates that of overall language usage – the Brown corpus at 1m words has considerably larger range than e.g. the Reuters corpus at 100m words.

The fact that psycholinguistic experiments need to control for word frequency has been known at least since Thorndike (1941) and frequency effects also play a key role in grammaticization (Bybee, 2003). Since the principal source of variability in word (n-gram) frequencies is the choice of topic, we can subsume overall considerations of genre under the selection of topics, especially as the former typically dictates the latter – for example, we rarely see literary prose or poetry dealing with undersea sedimentation rates. We assume a fixed inventory of topics $T_1, T_2, \ldots, T_k$, with k on the order 10^4, similar in granularity to the Northern Light topic hierarchy (Kornai et al 2003) and reserve T_0 to topicless texts or "General Language". Assuming that these topics appear in the language with frequency $q_1, q_2, \ldots, q_k$, summing to $1 - q_0 \leq 1$, the "average" topic is expected to have frequency about $1/k$ (and clearly, q_0 is on the same order, as it is very hard to find entirely topicless texts).

As is well known, the salience of different nouns and noun phrases appearing in the same structural position is greatly impacted not just by frequency (generally, less frequent words are more memorable) but also by stylistic value. For example, taboo words are more salient than neutral words of the same overall frequency. But style is also closely associated with topic, and if we match frequency profiles across topics we are therefore controlling for genre and style as well. Presenting psycholinguistical experiments is beyond the scope of this paper: here we put the emphasis on creating the computational resource, the frequency dictionary, that allows for detail matching of frequency profiles.

Defining the range r of a corpus C simply as $\sum_j q_j$ where the sum is taken over all topics touched by documents in C, single-author corpora typically have $r < 0.1$ even for encyclopedic writers, and web corpora have $r > 0.9$. Note that r just measures the range, it does *not* measure how representative a corpus is for some language community. Here we discuss results concerning all three ranges. For small range, we use the Hungarian translation of Orwell's *1984* – 98k words including punctuation tokens, (Dimitrova et al., 1998). For mid-range, we consider four topically segregated subcorpora of the Hungarian side of our Hungarian-English parallel corpus – 34m words, (Varga et al., 2005). For large-range we use our webcorpus – 700m words, (Halácsy et al., 2004).

1 Collecting and presenting the data

Hungarian lags behind "high density" languages like English and German but is hugely ahead of minority languages that have no significant machine readable material. Varga et al (2005) estimated there to be about 500 languages that fit in the same "medium density" category, together accounting for over 55% of the world's speakers. Halacsy et al (2004) described how a set of open source tools can be exploited to rapidly clean the results of web crawls to yield high quality monolingual corpora: the main steps are summarized below.

Raw data, preprocessing The raw dataset comes from crawling the top-level domain, e.g. `.hu`, `.cz`, `.hr`, `.pl` etc. Pages that contain no usable text are filtered out, and all text is converted to a uniform character encoding. Identical texts are dropped by checksum comparison of page bodies (a method that can handle near-identical pages, usually automatically generated, which differ only in their headers, datelines, menus, etc.)

Stratification A spellchecker is used to stratify pages by recognition error rates. For each page we measure the proportion of unrecognized (either incorrectly spelled or out of the vocabulary of the spellchecker) words. To filter out non-Hungarian (non-Czech, non-Croatian, non-Polish, etc.) documents, the threshold is set at 40%. If we lower the threshold to 8%, we also filter out *flat* native texts that employ Latin (7-bit) characters to denote their accented (8 bit) variants (these are still quite common due to the ubiquity of US keyboards). Finally, below the 4% threshold, webpages typically contain fewer typos than average printed documents, making the results comparable to older frequency counts based on traditional (printed) materials.

Lemmatization To turn a given stratum of the corpus into a frequency dictionary, one needs to collect the wordforms into lemmas based on the

same stem: we follow the usual lexicographic practice of treating inflected, but not derived, forms of a stem as belonging to the same lemma. Inflectional stems are computed by a morphological analyzer (MA), the choice between alternative morphological analyses is resolved using the output of a POS tagger (see Section 2 below). When there are several analyses that match the output of the tagger, we choose one with the least number of identified morphemes. For now, words outside the vocabulary of the MA are not lemmatized at all – this decision will be revisited once the planned extension of the MA to a morphological guesser is complete.

Topic classification Kornai et al (2003) presented a fully automated system for the classification of webpages according to topic. Combining this method with the methods described above enables the automatic creation of topic-specific frequency dictionaries and further, the creation of a per-topic frequency distribution for each lemma. This enables much finer control of word selection in psycholinguistic experiments than was hitherto possible.

1.1 How to present the data?

For Hungarian, the highest quality (4% threshold) stratum of the corpus contains 1.22m unique pages for a total of 699m tokens, already exceeding the 500m predicted in (Kilgarriff and Grefenstette, 2003). Since the web has grown considerably since the crawl (which took place in 2003), their estimate was clearly on the conservative side. Of the 699m tokens some 4.95m were outside the vocabulary of the MA (7% OOV in this mode, but less than 3% if numerals are excluded and the analysis of compounds is turned on). The remaining 649.7m tokens fall in 195k lemmas with an average 54 form types per lemma. If all stems are considered, the ratio is considerably lower, 33.6, but the average entropy of the inflectional distributions goes down only from 1.70 to 1.58 bits.

As far as the summary frequency list (which is less than a megabyte compressed) is concerned, this can be published trivially. Clearly, the availability of large-range gigaword corpora is in the best interest of all workers in language technology, and equally clearly, only open (freely downloadable) materials allow for replicability of experiments. While it is possible to exploit search engine queries for various NLP tasks (Lapata and Keller, 2004), for applications which use corpora as unsupervised training material downloadable base data is essential.

Therefore, a compiled webcorpus should contain actual texts. We believe all "cover your behind" efforts such as publishing only URLs to be fundamentally misguided. First, URLs age very rapidly: in any given year more than 10% become stale (Cho and Garcia-Molina, 2000), which makes any experiment conducted on such a basis effectively irreproducible. Second, by presenting a quality-filtered and characterset-normalized corpus the collectors actually perform a service to those who are less interested in such mundane issues. If everybody has to start their work from the ground up, many projects will exhaust their funding resources and allotted time before anything interesting could be done with the data. In contrast, the Free and Open Source Software (FOSS) model actively encourages researchers to reuse data.

In this regard, it is worth mentioning that during the crawls we always respected `robots.txt` and in the two years since the publication of the gigaword Hungarian web corpus, there has not been a single request by copyright holders to remove material. We do not advocate piracy: to the contrary, it is our intended policy to comply with removal requests from copyright holders, analogous to Google cache removal requests. Finally, even with copyright material, there are easy methods for preserving interesting linguistic data (say unigram and bigram models) without violating the interests of businesses involved in selling the running texts. [1]

2 The disambiguation of morphological analyses

In any morphologically complex language, the MA component will often return more than one possible analysis. In order to create a lemmatized frequency dictionary it is necessary to decide which MA alternative is the correct one, and in the vast majority of cases the context provides sufficient information for this. This morphological disambiguation task is closely related to, but not identical with, part of speech (POS) tagging, a term we reserve here for finding the major parts

[1] This year, we are publishing smaller pilot corpora for Czech (10m words), Croatian (4m words), and Polish (12m words), and we feel confident in predicting that these will face as little actual opposition from copyright holders as the Hungarian Webcorpus has.

of speech (N, V, A, etc). A full tag contains both POS information and morphological annotation: in highly inflecting languages the latter can lead to tagsets of high cardinality (Tufiş et al., 2000). Hungarian is particularly challenging in this regard, both because the number of ambiguous tokens is high (reaching 50% in the Szeged Corpus according to (Csendes et al., 2004) who use a different MA), and because the ratio of tokens that are not seen during training (unseen) can be as much as four times higher than in comparable size English corpora. But if larger training corpora are available, significant disambiguation is possible: with a 1 m word training corpus (Csendes et al., 2004) the TnT (Brants, 2000) architecture can achieve 97.42% overall precision.

The ratio of ambiguous tokens is usually calculated based on alternatives offered by a morphological lexicon (either built during the training process or furnished by an external application; see below). If the lexicon offers alternative analyses, the token is taken as ambiguous irrespective of the probability of the alternatives. If an external resource is used in the form of a morphological analyzer (MA), this will almost always overgenerate, yielding false ambiguity. But even if the MA is tight, a considerable proportion of ambiguous tokens will come from legitimate but rare analyses of frequent types (Church, 1988). For example the word *nem*, can mean both 'not' and 'gender', so both `ADV` and `NOUN` are valid analyses, but the adverbial reading is about five orders of magnitude more frequent than the noun reading, (12596 vs. 4 tokens in the 1 m word manually annotated Szeged Korpusz (Csendes et al., 2004)).

Thus the difficulty of the task is better measured by the average information required for disambiguating a token. If word w is assigned the label T_i with probability $P(T_i|w)$ (estimated as $C(T_i, w)/C(w)$ from a labeled corpus) then the label entropy for a word can be calculated as $H(w) = -\sum_i P(T_i|w) \log P(T_i|w)$, and the difficulty of the labeling task as a whole is the weighted average of these entropies with respect to the frequencies of words w: $\sum_w P(w)H(w)$. As we shall see in Section 3, according to this measure the disambiguation task is not as difficult as generally assumed.

A more persistent problem is that the ratio of unseen items has very significant influence on the performance of the disambiguation system. The problem is more significant with smaller corpora: in general, if the training corpus has N tokens and the test corpus is a constant fraction of this, say $N/10$, we expect the proportion of new words to be cN^{q-1}, where q is the reciprocal of the Zipf constant (Kornai, 1999). But if the test/train ratio is not kept constant because the training corpus is limited (manual tagging is expensive), the number of tokens that are not seen during training can grow very large. Using the 1.2 m words of Szeged Corpus for training, in the 699 m word webcorpus over 4% of the non-numeric tokens will be unseen. Given that TnT performs rather dismally on unseen items (Oravecz and Dienes, 2002) it was clear from the outset that for lemmatizing the webcorpus we needed something more elaborate.

The standard solution to constrain the probabilistic tagging model for some of the unseen items is the application of MA (Hakkani-Tür et al., 2000; Hajič et al., 2001; Smith et al., 2005). Here a distinction must be made between those items that are not found in the training corpus (these we have called *unseen* tokens) and those that are not known to the MA – we call these out of vocabulary (OOV). As we shall see shortly, the key to the best tagging architecture we found was to follow different strategies in the lemmatization and morphological disambiguation of OOV and known (invocabulary) tokens.

The first step in tagging is the annotation of inflectional features, with lemmatization being postponed to later processing as in (Erjavec and Džeroski, 2004). This differs from the method of (Hakkani-Tür et al., 2000), where all syntactically relevant features (including the stem or lemma) of word forms are determined in one pass. In our experience, the choice of stem depends so heavily on the type of linguistic information that later processing will need that it cannot be resolved in full generality at the morphosyntactic level.

Our first model (MA-ME) is based on disambiguating the MA output in the maximum entropy (ME) framework (Ratnaparkhi, 1996). In addition to the MA output, we use ME features coding the surface form of the preceding/following word, capitalization information, and different character length suffix strings of the current word. The MA used is the open-source `hunmorph` analyzer (Trón et al., 2005) with the `morphdb.hu` Hungarian morphological resource, the ME is the OpenNLP package (Baldridge et al., 2001). The

MA-ME model achieves 97.72% correct POS tagging and morphological analysis on the test corpus (not used in training).

Maximum entropy or other discriminative Markov models (McCallum et al., 2000) suffer from the label bias problem (Lafferty et al., 2001), while generative models (most notably HMMs) need strict independence assumptions to make the task of sequential data labeling tractable. Consequently, long distance dependencies and non-independent features cannot be handled. To cope with these problems we designed a hybrid architecture, in which a trigram HMM is combined with the MA in such a way that for tokens known to the MA only the set of possible analyses are allowed as states in the HMM whereas for OOVs all states are possible. Lexical probabilities $P(w_i|t_i)$ for seen words are estimated from the training corpus, while for unseen tokens they are provided by the the above MA-ME model. This yields a trigram HMM where emission probabilities are estimated by a weighted MA, hence the model is called WMA-T3. This improves the score to 97.93%.

Finally, it is possible to define another architecture, somewhat similar to Maximum Entropy Markov Models, (McCallum et al., 2000), using the above components. Here states are also the set of analyses the MA allows for known tokens and all analyses for OOVs, while emission probabilities are estimated by the MA-ME model. In the first pass TnT is run with default settings over the data sequence, and in the second pass the ME receives as features the TnT label of the preceding/following token as well as the one to be analyzed. This combined system (TnT-MA-ME) incorporates the benefits of all the submodules and reaches an accuracy of 98.17% on the Szeged Corpus. The results are summarized in Table 1.

model	accuracy
TnT	97.42
MA+ME	97.72
WMA+T3	97.93
TnT+MA+ME	98.17

Table 1: accuracy of morphological disambiguation

We do not consider these results to be final: clearly, further enhancements are possible e.g. by a Viterbi search on alternative sentence taggings using the T3 trigram tag model or by handling OOVs on a par with known unseen words using the guesser function of our MA. But, as we discuss in more detail in Halacsy et al 2005, we are already ahead of the results published elsewhere, especially as these tend to rely on idealized MA systems that have their morphological resources extended so as to have no OOV on the test set.

3 Conclusions

Once the disambiguation of morphological analyses is under control, lemmatization itself is a mechanical task which we perform in a database framework. This has the advantage that it supports a rich set of query primitives, so that we can easily find e.g. nouns with back vowels that show stem vowel elision and have approximately the same frequency as the stem *orvos* 'doctor'. Such a database has obvious applications both in psycholinguistic experiments (which was one of the design goals) and in settling questions of theoretical morphology. But there are always nagging doubts about the closed world assumption behind databases, famously exposed in linguistics by Chomsky's example *colorless green ideas sleep furiously:* how do we distinguish this from **green sleep colorless furiously ideas* if the observed frequency is zero for both?

Clearly, a naive empirical model that assigns zero probability to each unseen word form makes the wrong predictions. Better estimates can be achieved if unseen words which are known to be possible morphologically complex forms of seen lemmas are assigned positive probability. This can be done if the probability of a complex form is in some way predictable from the probabilities of its component parts. A simple variant of this model is the positional independence hypothesis which takes the probabilities of morphemes in separate positional classes to be independent of each other. Here we follow Antal (1961) and Kornai (1992) in establishing three positional classes in the inflectional paradigm of Hungarian nouns.

```
# Position 1 parameters
FAM                     0.0001038986
PLUR                    0.1372398793
PLUR_POSS               0.0210927964
PLUR_POSS<1>            0.0011609442
PLUR_POSS<1><PLUR>      0.0028751247
PLUR_POSS<2>            0.0004958278
PLUR_POSS<2><PLUR>      0.0000740203
PLUR_POSS<PLUR>         0.0023850120
POSS                    0.1461635946
```

POSS<1>	0.0073305415
POSS<1><PLUR>	0.0073652648
POSS<1>_FAM	0.0000092294
POSS<2>	0.0027628071
POSS<2><PLUR>	0.0003006440
POSS<2>_FAM	0.0000030591
POSS<PLUR>	0.0069613929
POSS_FAM	0.0000000001
ZERO1	0.6636759634
# Position 2 parameters	
ANP	0.0007780001
ANP<PLUR>	0.0000248301
ZERO2	0.9991971698
# Position 3 parameters	
CAS<ABL>	0.0078638013
CAS<ACC>	0.1346412632
CAS<ADE>	0.0045132704
CAS<ALL>	0.0138677701
CAS<CAU>	0.0037332025
CAS<DAT>	0.0301123636
CAS<DEL>	0.0128222999
CAS<ELA>	0.0118596792
CAS<ESS>	0.0010230505
CAS<FOR>	0.0031204983
CAS<ILL>	0.0154186683
CAS<INE>	0.0582887516
CAS<INS>	0.0406197868
CAS<SBL>	0.0386519707
CAS<SUE>	0.0357416253
CAS<TEM>	0.0013095685
CAS<TER>	0.0034032438
CAS<TRA>	0.0017860054
ZERO3	0.5812231804

Table 3: marginal probabilities in noun inflection

The innermost class is used for number and possessive, with a total of 18 choices including the zero morpheme (no possessor and singular). The second positional class is for anaphoric possessives with a total of three choices including the zero morpheme, and the third (outermost) class is for case endings with a total of 19 choices including the zero morpheme (nominative) for a total of 1026 paradigmatic forms. The parameters were obtained by downhill simplex minimization of absolute errors. The average absolute error is of the values computed by the independece hypothesis from the observed values is 0.000099 (mean squared error is $9.18 \cdot 10^{-7}$), including the 209 paradigmatic slots for which no forms were found in the webcorpus at all (but the independence model will assign positive probability to any of them as the product of the component probabilities). When checking the independence hypothesis with Φ statistics in the webcorpus for every nominal inflectional morpheme pair the members of which are from different dimensions, the Φ coefficient remained less than 0.1 for each pair but 3. For these 3 the coefficient is under 0.2 (which means that the shared variance of these pairs is between 1% and 2%) so we have no reason to discard the independence hypothesis. If we run the same test on the 150 million words Hungarian National Corpus, which was analyzed and tagged by different tools, we also get the same result (Nagy, 2005).

It is very easy to construct low probability combinations using this model. Taking a less frequent possessive ending such as the 2nd singular posessor familiar plural *-odék*, the anaphoric plural *-éi*, and a rarer case ending such as the formalis *-ként* we obtain combinations such as *barátodékéiként* "as the objects owned by your friends' company". The model predicts we need a corpus with about $4.2 \cdot 10^{12}$ noun tokens to see this suffix combination (not necessarily with the stem *barát* "friend") or about ten trillion tokens. While the current corpus falls short by four orders of magnitude, this is about the contribution of the anaphoric plural (which we expect to see only once in about 40k noun tokens) so for any two of the three position classes combined the prediction that valid inflectional combinations will actually be attested is already testable.

Using the fitted distribution of the position classes, the entropy of the nominal paradigm is computed simply as the sum of the class entropies, 1.554 + 0.0096 + 2.325 or 3.888 bits. Since the nominal paradigm is considerably more complex than the verbal paradigm (which has a total of 52 forms) or the infinitival paradigm (7 forms), this value can serve as an upper bound on the inflectional entropy of Hungarian. In Table 3 we present the actual values, computed on a variety of frequency dictionaries. The smallest of these is based on a single text, the Hungarian translation of Orwell's *1984*. The mid-range corpora used in this comparison are segregated in broad topics: law (EU laws and regulations), literature, movie subtitles, and software manuals: all were collected from the web as part of building a bilingual English-Hungarian corpus. Finally, the large-range is the full webcorpus at the best (4% reject) quality stratum.

	1984	law	literature	subtitles	software	webcorpus
token	98292	2310742	7971157	2667420	839339	69926550
type	20343	110040	431615	188131	81729	2083023
OOV token	3141	266368	335660	181292	140551	4951743
OOV type	1132	39467	87574	50078	45799	994890
lemma	10644	60602	165259	85491	58939	1189471
lemma excl. OOV	9513	21136	77686	35414	13141	194589
lemma entropy	1.14282	1.04118	1.54922	1.41374	1.14516	1.57708
lemma entropy excl. OOV	1.18071	1.17687	1.61753	1.51718	1.37559	1.69743

Table 3: inflectional entropy of Hungarian computed on a variety of frequency dictionaries

Our overall conclusion is that for many purposes a web-based corpus has significant advantages over more traditional corpora. First, it is cheap to collect. Second, it is sufficiently heterogeneous to ensure that language models based on it generalize better on new texts of arbitrary topics than models built on (balanced) manual corpora. As we have shown, automatically tagged and lemmatized webcorpora can be used to obtain large coverage stem and wordform frequency dictionaries. While there is a significant portion of OOV entries (about 3% for our current MA), in the design of psycholinguistic experiments it is generally sufficient to consider stems already known to the MA, and the variety of these (over three times the stem lexicon of the standard Hungarian frequency dictionary) enables many controlled experiments hitherto impossible.

References

László Antal. 1961. A magyar esetrendszer. *Nyelvtudományi Értekezések*, 29.

Jason Baldridge, Thomas Morton, and Gann Bierner. 2001. The opennlp maximum entropy package. *http://maxent.sourceforge.net*.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*, Seattle, WA.

Joan Bybee. 2003. Mechanisms of change in grammaticization: the role of frequency. In Brian Joseph and Richard Janda, editors, *Handbook of Historical Linguistics*, pages 602–623. Blackwell.

Junghoo Cho and Hector Garcia-Molina. 2000. The evolution of the web and implications for an incremental crawler. In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pages 200–209, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kenneth Ward Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the second conference on Applied natural language processing*, pages 136–143, Morristown, NJ, USA. Association for Computational Linguistics.

Dóra Csendes, János Csirik, and Tibor Gyimóthy. 2004. The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. In Karel Pala Petr Sojka, Ivan Kopecek, editor, *Text, Speech and Dialogue: 7th International Conference, TSD*, pages 41–47.

Ludmila Dimitrova, Tomaz Erjavec, Nancy Ide, Heiki Jaan Kaalep, Vladimir Petkevic, and Dan Tufiş. 1998. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern european languages. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 315–319, San Francisco, California. Morgan Kaufmann Publishers.

Tomaž Erjavec and Sašo Džeroski. 2004. Machine learning of morphosyntactic structure: Lemmatizing unknown Slovene words. *Applied Artificial Intelligence*, 18(1):17–41.

Jan Hajič, Pavel Krbec, Karel Oliva, Pavel Květoň, and Vladimír Petkevič. 2001. Serial combination of rules and statistics: A case study in Czech tagging. In *Proceedings of the 39th Association of Computational Linguistics Conference*, pages 260–267, Toulouse, France.

Dilek Z. Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. 2000. Statistical morphological disambiguation for agglutinative languages. In *Proceedings of the 18th conference on Computational linguistics*, pages 285–291, Morristown, NJ, USA. Association for Computational Linguistics.

Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. 2004. Creating open language resources for Hungarian. In *Proceedings of Language Resources and Evaluation Conference (LREC04)*. European Language Resources Association.

Péter Halácsy, András Kornai, and Dániel Varga. 2005. Morfológiai egyértelműsítés maximum entrópia módszerrel (morphological disambiguation with the maxent method). In *Proc. 3rd Hungarian Computational Linguistics Conf.* Szegedi Tudományegyetem.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–348.

András Kornai, Marc Krellenstein, Michael Mulligan, David Twomey, Fruzsina Veress, and Alec Wysoker. 2003. Classifying the hungarian web. In A. Copestake and J. Hajic, editors, *Proc. EACL*, pages 203–210.

András Kornai. 1992. Frequency in morphology. In I. Kenesei, editor, *Approaches to Hungarian*, volume IV, pages 246–268.

András Kornai. 1999. Zipf's law outside the middle range. In J. Rogers, editor, *Proc. Sixth Meeting on Mathematics of Language*, pages 347–356. University of Central Florida.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

Mirella Lapata and Frank Keller. 2004. The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 121–128, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Anke Luedeling, Stefan Evert, and Marco Baroni. 2005. Using web data for linguistic purposes. In Marianne Hundt, Caroline Biewer, and Nadjia Nesselhauf, editors, *Corpus linguistics and the Web*. Rodopi.

Andrew McCallum, Dayne Freitag, and Fernando Pereira. 2000. Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of the 17th International Conference on Machine Learning*, pages 591–598. Morgan Kaufmann, San Francisco, CA.

Viktor Nagy. 2005. A magyar főnévi inflexió statisztikai modellje (statistical model of nominal inflection in hungarian. In *Proc. Kodolányi-ELTE Conf.*

Csaba Oravecz and Péter Dienes. 2002. Efficient stochastic part-of-speech tagging for Hungarian. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC2002)*, pages 710–717.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In Karel Pala Petr Sojka, Ivan Kopecek, editor, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142, University of Pennsylvania.

Noah A. Smith, David A. Smith, and Roy W. Tromble. 2005. Context-based morphological disambiguation with random fields. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver.

Edward L. Thorndike. 1941. *The Teaching of English Suffixes*. Teachers College, Columbia University.

Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. 2005. Hunmorph: open source word analysis. In *Proceeding of the ACL 2005 Workshop on Software*.

Dan Tufiş, Péter Dienes, Csaba Oravecz, and Tamás Váradi. 2000. Principled hidden tagset design for tiered tagging of Hungarian. In *Proceedings of the Second International Conference on Language Resources and Evaluation*.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing 2005 Conference*, pages 590–596, Borovets. Bulgaria.