Creating open language resources for Hungarian

Péter Halácsy*, **András Kornai**†, **László Németh***, **András Rung** *, **István Szakadát***, **Viktor Trón**‡

*Budapest Institute of Technology Media Research and Education Center
{halacsy,nemeth,rung,szakadat}@mokk.bme.hu
†MetaCarta Inc., andras@kornai.com
‡International Graduate College, Saarland University and University of Edinburgh, v.tron@ed.ac.uk

Abstract

The paper provides an overview of the open source Hungarian language resources that the *SzóSzablya* 'WordSword' project is creating. An extensive crawl of the .hu domain yielded a raw dataset of over 18m web pages. We discuss the methods used to detect and remove duplicates, low quality, foreign, and mixed language documents, and describe the resulting gigaword corpus and various frequency counts and dictionaries based on it.

1. Introduction

With Hungary's ascension to the EU, wider availability of Hungarian language resources (LRs) is becoming more critical. Various Hungarian LRs such as corpora, word lists, frequency counts, and machine readable dictionaries already exist, as do language technology tools (LTs) such as tokenizers, stemmers, spellcheckers, morphological analyzers, POS taggers etc.[1] These are, however, for the most part proprietary products: the companies and research labs developing them are often reluctant to make them available even for research, let alone commercial purposes.

The *SzóSzablya* 'WordSword' project at the Centre of Media Research and Education of Budapest University of Technology and Economics started in March 2003 with the express goal to offer a solution to this problem by developing a comprehensive set of LRs with an LT toolkit which are made publicly available under an unrestrictive LGPL-style license. The body of this paper is organized as follows. Section 2 describes the process of creating the gigaword web2 corpus, the project's major resource, focusing on the methods used for collecting and cleaning the data. Section 3 discusses the frequency counts and dictionaries that have been compiled on the basis of this corpus. Section 4 concludes by sketching future directions of the project.

2. The Hungarian Web Corpus

In a pilot study the Axelero web crawler was used to collect approximately six million web pages from the .hu domain. Duplicate pages were detected by identical MD5 checksums, and documents were stripped of HTML tags. Tokenization was performed by breaking on punctuation, hyphens and whitespace, and the resulting tokens were uppercased. This resulted in a corpus of over 2 billion word tokens. Document frequency (DF) counts for words and word pairs were calculated yielding 31.1 million unigram types out of which 18.3 million were DF hapaxes.[2]

A series of experiments pilot0, pilot1, web0, and web1 helped us refine our methodology. First, we created a more sophisticated duplicate detection algorithm that will also eliminate duplicate pages that differ only in irrelevant detail such as auto-generated dates or headers. Second, we concluded that the initial text normalization and tokenization methods obscured a great deal of valuable detail, and switched to case preservation and a more complex tokenization scheme. Third, we found that in n-gram counts, text frequency (TF) numbers are more useful than DF numbers, and changed our infrastructure accordingly. Fourth, and perhaps most important, we succeeded in identifying the major sources of noise in the data (non-Hungarian language pages and raw file formats such as pdf, doc, mime64 etc.) and developed a tunable filtering step to remove these. Here we omit the evolutionary details, and concentrate on the current version of the methods used in creating the web2 gigaword corpus and attendant frequency counts that *SzóSzablya* is making public.

The web2 corpus gathered in the main study is based on 18m pages, and takes up over 50GB compressed.[3] As a comparison, the Hungarian National Corpus[4] (Váradi 2002) is 153.7m words (300MB compressed), the Hungarian Historical Corpus[5] (Pajzs 2000) is 24.5m words (50MB compressed), the Szeged Corpus[6] (Alexin etal 2003) is 1m words (8MB compressed), the machine-readable version of Orwell's 1984 created for the Copernicus project (Erjavec and Ide 1998)[7] is 81k words (220k compressed). These corpora are all considerably smaller than our present collection, and are not available for commercial research and development.[8]

Raw data set sizes do not provide an adequate basis for comparison, however. By the time duplicate pages and obviously non-Hungarian documents are disposed of and HTML markup is stripped, crawl-based corpora can shrink by an order of magnitude. As we shall see in Section 3,

[1] For a synopsis and a non-exhaustive listing of resources, see the project website www.szoszablya.hu

[2] The pilot0 DF count is also made publicly available courtesy of Axelero Internet.

[3] The entire raw data is available on request. Smaller datasets are available through anonymous ftp (ftp.szoszablya.hu).

[4] corpus.nytud.hu/mnsz/index_eng.html

[5] www.nytud.hu/hhc

[6] www.inf.u-szeged.hu/lll/szegedcorpus.html

[7] corpus.nytud.hu/demo/infotrend/orwell

[8] To our knowledge, only the SZTAKI corpus (also based on a webcrawl, 2.6m web pages before duplicate elimination, 8GB compressed) is of comparable size. This LR is also made publicly available from the project repository courtesy of SZTAKI.

the main factor affecting further deflation is the stringency of the selection criteria used to ensure the quality of the data. Since web content is quite diverse in terms of both genre and compliance with norms, the quality of the data is much harder to guarantee than in the case of texts from controlled sources such as newspapers or edited prose. This makes the comparison of data sizes difficult, and the matter is further complicated by the added value of linguistic information, such as morphological analysis or word sense annotation, which depends greatly on whether the results are machine-generated or hand-corrected (all the corpora mentioned above contain annotation and are to varying degrees also manually disambiguated). In order to create a corpus of Hungarian texts of reasonable quality, the raw data set needs to be cleaned. This involves several filtering steps to which we now turn.

For normalization we use HunNorm, which performs HTML stripping and character conversion to produce uniform text files from web pages. It uses a `flex` pipeline and relies on existing open source code such as `GNU Recode` for UTF-8 conversion and `file` for determining file types and removing binary files. HunNorm typically deflates the results by 50% or more.

Next we detect sentence boundaries by the HunToken module, a rule based tokenizer written in `flex` which is similar in concept and design to the rule system described Mikheev (2002). It employs 25 regular-expression rules, and relies on an approximately 150-word list of common abbreviations. Evaluated against the Szeged Corpus, HunToken's sentence boundaries are incorrect in 1064 cases out of the 86094 sentences, yielding an error rate of 1.3% which is significantly better than the simple regex [.!?] baseline of 6083 (7.0%).

By establishing sentence boundaries we can take into account that script-generated text (such as headlines, dates, tables of content) are typically not part of ordinary sentence structure. If we eliminate all extrasentential material and compute checksums based on the sentence bodies alone, we can detect script-generated variants of the same page and eliminate linguistically empty pages. The similarity method suggested in Chakrabarti (2001) is capable of detecting block-edited/paraphrased variants as well: our method is not as sensitive but considerably less intensive computationally. This step alone deflates the corpus by more than 50%: the resulting `web2`, 3.5m pages, is smaller than the raw pilot, but incomparably better quality.

3. The frequency dictionary

Since existing corpora for Hungarian are not available or downloadable, even basic frequency counts for arbitrary units such as n-grams or letters are impossible to obtain. Individual DF values from Hungarian Historical Corpus can be obtained through a web interface, but to this day the only publicly available batch resource for word frequency counts in Hungarian is Füredi and Kelemen's (1989) frequency dictionary (henceforth FK89), based on a 500k word belles lettres corpus.[9]

While `web2` is a significant LR in itself e.g. for statistical n-gram modelling, most applications require better selected and more thoroughly processed data, such as provided by a *frequency dictionary* where morphologically related entries are collected in the same lemma, and, ideally, homonyms such as nap_1 'sun' and nap_2 'day' are separated. One of our major objectives is to develop such a dictionary, based on a corpus three orders of magnitude larger, and encompassing more than just literary usage.

In general, the most important decisions on frequency counts are the ones made earliest: in addition to corpus selection, we call special attention to the tokenization step. To see how large impact low-level tokenization decisions can have on the absolute and relative frequency values, in table 1 we compare the top 20 entries from `pilot0`, which uses a primitive regex $[.!?]\backslash s$ tokenizer and upcasing, to the top 20 from `web2`, which uses the more sophisticated HunToken algorithm.

`pilot0`		`web2`	
HU	4516525	a	2702036
A	3479829	és	2368346
LISTS	3411785	az	2300925
DIRECTORIES	3406266	A*	2228939
AZ	2432533	is	1827309
ÉS	2210614	nem	1678326
IS	1959822	hogy	1657968
1	1774391	Az*	1624776
E	1633924	egy	1573182
NEM	1631758	meg	1378270
2	1574935	csak	1159372
HTML	1568672	van	1124243
VAN	1518679	de	1113425
EZ	1479599	vagy	1107128
HOGY	1472649	már	1035983
EGY	1445847	el	1027588
3	1326171	még	981011
2001	1310325	ki	902715
10	1278561	mint	892048
MEG	1270426	ha	885077

Table 1: The top 20 unigram DF values in the pilot and main studies

As the table shows rather strikingly, minor changes in tokenization, such as separating the components of URLs in the pilot, but not in the main count, will radically alter the ranking. *hu,* an emphatic particle of Hungarian, does not even make it to the top 100k once it is kept distinct from the `.hu` domain name suffix. HunToken recognizes categories like punctuation, numbers, date and time formats etc.[10]

Since HunToken also provides sentence-level chunking, we can preserve a great deal of positional information

[9]Until recently, only the top few thousand lemmas of FK89 were available in hardcopy, though simplified frequency data from FK89 could be obtained from the widely used SZÓTÁR lexical database (Füredi, Kornai, and Prószéky 2004). Both FK and SZÓTÁR are now available in our repository (`www.szoszablya.hu`) courtesy of their authors.

[10]Our token classification follows that of the Szeged Corpus, which utilizes extended TEI LITE XML document format with MSD morphological codes.

about tokens, thereby enabling simple (n-gram free) disambiguation strategies in subsequent lemmatization steps. For example, sentence initial occurrences can be treated as separate tokens (marked by an appended asterisk): this is especially useful in distinguishing proper names and homonymous common nouns. For example, *Kovács* ('Smith', the most common Hungarian family name) occurs 88307 times medially while *kovács* 'blacksmith' occurs only 2785 times. Sentence-initially, where the two senses appear as the ambiguous *Kovács*, it occurs 28667 times. Frequencies of the ambiguous senses can then be estimated on the basis of the non-ambiguous occurrences, which is correct if the position in question is independent of the sense.

The raw data set for `web2` is about 18.7m pages (50GB compressed). After the removal of executables and other non-textual pages, the elimination of HTML markup, and duplicate page removal, the actual `web2` corpus is about 3.5m documents (5.2GB compressed), including many foreign and mixed language documents. Compared to literary or journalistic prose the quality of this material is very uneven: there is a great deal of computer jargon, telegraphic SMS- and chat-speak, and a considerable number of *flat* pages (Kornai and Tóth 1997) which replace some Hungarian accented characters by their 7-bit ascii counterparts. While the *SzóSzablya* project did not wish to pass normative judgements on such pages, it was clear from the outset that for many applications it is desirable to stratify the corpus by some measure of 'correctness', and we chose adherence to official Hungarian spelling (a matter very closely regulated by the Hungarian Academy of Sciences) as our yardstick. We run every document through a spellchecker, and in stratified subcorpora retain only pages that contain no more than t% spelling errors.

The spellchecker we use is HunSpell, also a module of our open source LT toolkit. HunSpell uses an ispell derivative, the extended version of OpenOffice.org's MySpell spell checking library and is historically the earliest tool at our disposal. Many improvements in HunSpell became part of the original MySpell library. The spellchecker itself is language independent, the resource files we used for Hungarian are all open source and provide excellent Hungarian spellchecking (for a comparison with the market-leading closed source spellchecker, see Németh 2003).

Setting t to 40 can reliably filter out non-Hungarian documents while keeping even extremely low-quality (e.g. flat) Hungarian pages. Setting t to 8 will also eliminate flat pages, but retains geek jargon and other non-standard text. Setting t to 4 leaves only documents that have fewer typos than average printed materials. Table 2 shows the major parameters of the corpus strata (t=100 corresponds to no spelling-based filtering):

t (%)	100	40	8	4
pages (m)	3.493	3.125	1.918	1.221
tokens (m)	1486	1310	928	589
types (m)	19.1	15.4	10.9	7.2
hapaxes (m)	11.5	8.9	6.3	4.2

Table 2: Stratified corpus size

The frequency distribution of spelling error percentages in `web2` has a strongly bimodal profile: many pages have very few errors, many pages have many errors, but only a few pages exist with about half of their text spelled incorrectly. Manual checking makes clear that documents with many spelling errors are predominantly foreign language pages, where correctly spelled Hungarian words can only result from direct quotations, proper names, and homographic vocabulary items such as Hungarian *fuss* 'run' vs German *fuss* 'foot' vs English *fuss* 'id'. There are plenty of orthographically unassimilated loans like *standard, project* (though over time these tend to be replaced by their assimilated counterparts *sztenderd, projekt*), and there are some etymologically related items, but on the whole Hungarian is sufficiently dissimilar to other languages to make the spellchecker based method a surprisingly reliable language identification tool. To see this, consider the document frequencies of the Hungarian definite article *a/az* and the English definite article *the* in table 3. Manual sampling of the remaining instances of *the* makes clear that they appear in high-quality documents, e.g., Hungarian language newspapers mentioning *The Times*.

t (%)	100	40	8	4
The*	143	30	6	2
The	131	94	27	12
the	333	156	38	14
Az*	2033	2169	2094	2086
Az	305	323	311	301
az	2884	3072	2899	2844

Table 3: Stratified DF of definite articles

While we consider the gigaword stratum (928m words in the documents with less than 8% spellcheck error) to be quite representative of contemporary Hungarian usage, to obtain results more comparable to FK89 we also consider the higher quality $t = 4$ stratum (589m words). But because genre is a strong predictor of frequency, the data in FK89 does not correlate well with our results at any cutoff (Pearson's c=0.64 for log frequencies of words that appear in both samples, while the strata correlate with each other at 0.98 or better), and we believe that in spite of its smaller sample size FK89 reflects actual usage frequencies in the literary domain more reliably than *web2*. But to the extent that the web is more representative of a person's inventory of genres, for many purposes ranging from spellchecking to psycholinguistic research, the web could provide a better frequency model.

By collapsing words with the same stem into one lemma, we obtain an *approximate* frequency dictionary (only approximate, because at this stage neither stemming ambiguities nor homonyms are resolved). Lemmatization was performed by HunStem, which is an extended version of the HunSpell library, following the same affix stripping rules. In addition to providing a stem (or, in case of ambiguity, multiple stem candidates), HunStem also outputs partial morphological analysis information, which makes it possible to correctly lemmatize exceptions. The top 15 lemmas with the relevant counts are shown in table 4.

stem	$t = 100$		$t = 40$		$t = 8$		$t = 4$	
	forms	tf	forms	tf	forms	tf	forms	tf
a	1	112413828	1	109118173	1	80666377	1	52769698
az	68	47064698	68	46562937	68	34898956	67	23155708
és	1	27035824	1	26847070	1	19862073	1	12726963
van	138	23794027	136	23395869	126	16364903	115	10157192
hogy	1	16585835	1	16407853	1	12106465	1	7781361
nem	153	15956745	153	15714855	146	11119096	128	6863047
is	1	15824358	1	14300654	1	10109707	1	6290339
ez	53	11846524	53	11694109	48	8631668	43	5616677
egy	79	11438348	79	11287625	67	7756493	58	4536819
meg	1	6529862	1	6415950	1	4421274	1	2798180
de	1	6414373	1	5808632	1	3856245	1	2230653
ha	1	5648497	1	5541838	1	3893474	1	2467018
csak	1	5080107	1	5005367	1	3469396	1	2099715
kell	66	4556123	66	4492710	63	3436836	56	2392951
már	1	4119406	1	4101918	1	2905754	1	1725669

Table 4: Number of forms and frequencies for the 15 most frequent lemmas

The approximate lemmatization used in this table collapses sentence-initial with non-initial variants, and collapses case distinctions present in the original text. While the list is dominated by indeclinabilia, some words, in particular the copula *van* 'be' and the demonstrative *az* 'that' have many affixed forms which boost its rank considerably compared to table 1, which reflects only the zero affixed (3rd person singular present) copular form.

4. Future directions

Our next obvious step toward a full frequency dictionary is to replace the approximate (stemming-based) lemmatization used so far by a more precise morphological analysis. We have already created a prototype morphological analyzer, HunMorph, using the same open libraries, but incorporating substantial extensions to the underlying ispell analysis such as the ability to return multiple morphological parses of ambiguous forms and the possibility to handle homonymous stems. Most importantly, HunMorph allows a two-stage process of suffix stripping, whereby it can trade its efficiency to overcome memory limitations resulting from productive suffix-combinations.

To improve the stem dictionary and the morphological grammar, we are also developing an off-line preprocessor HunLex that supplies the analysis tools with configured lexical resource files by compiling HunSpell-style dictionary and affix files.

This paper discussed our first steps in creating LRs for Hungarian. Some modules of our LT toolkit are discussed in a companion paper (Németh et al. 2004), but this paper focused on the process of creating a gigaword corpus from scratch. Given that gigaword corpora currently exist only for a handful of languages and are greatly copyright-encumbered, our methods may be of general interest.

Acknowledgements

The SzóSzablya project is funded by an ITEM grant from the Hungarian Ministry of Informatics and Telecommunications, and benefits greatly from logistic and infrastructural support of MATÁV Rt. and Axelero Internet. Special thanks to Gábor Kiss (Axelero).

5. References

Alexin, Z., J. Csirik, T. Gyimóthy, K. Bibok, Cs. Hatvani, G. Prószéky and L. Tihanyi (2003). Manually annotated Hungarian Corpus. Proc. of Research Note Sessions of 10th EACL. Budapest. 53–56.

Chakrabarti, S. (2003). *Mining the web*. Morgan Kaufmann.

Erjavec, T. and N. Ide (1998): The MULTEXT-East Corpus. Proc. of LREC'98.

Füredi, M. and J. Kelemen. (1989). A mai magyar nyelv széppróza i gyakorisági szótára. [Frequency dictionary of present day literary Hungarian]. Akadémiai. Budapest.

Füredi, M., A. Kornai and G. Prószéky G. (2004): The SZÓTÁR database. (In Hungarian). ms. URL `http://www.szoszablya.hu/`.

Halácsy, P., A. Kornai, L. Németh, A. Rung, I. Szakadát and V. Trón (2003). Szógyakoriság és helyesírás-ellenőrzés. [Word frequency and spell-checker accuracy] (Hung. with English summary). Proc. of the 1st Hungarian Computational Linguistics Conference. 211–217.

Kornai, A. and G. Tóth (1997). Computer generation of accent marks. (Hung. with English summary) Magyar Tudomány 1997/4. 400–410.

Mikheev, A. (2002). Periods, Capitalized Words, etc. *Computational Linguistics* **28**:289–318.

Németh, L. (2003). A Szószablya fejlesztés. 5th Hungarian Linux Conference. URL `http://konf2003.linux.hu/`.

Németh, L., P. Halácsy, A. Kornai, L., A. Rung, I. Szakadát and V. Trón (2004). A stemmer based on ispell technology. To be presented at SALTMIL 2004.

Pajzs, J. (2000): Making Historical Dictionaries with the Computer. Proc. of EURALEX 2000. Stuttgart. 249–259.

Váradi, T. (2002) The Hungarian National Corpus. LREC 2002. 385–389.