Parallel corpora for medium density languages

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

Péter Halácsy
MREC
Stoczek u. 2
H-1111 Budapest
halacsy@mokk.bme.hu

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

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

László Németh
MREC
Stoczek u. 2
H-1111 Budapest
nemeth@mokk.bme.hu

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

Abstract

The choice of natural language technology appropriate for a given language is greatly impacted by *density* (availability of digitally stored material). More than half of the world speaks medium density languages, yet many of the methods appropriate for high or low density languages yield suboptimal results when applied to the medium density case. In this paper we describe a general methodology for rapidly collecting, building, and aligning parallel corpora for medium density languages, illustrating our main points on the case of Hungarian, Romanian, and Slovenian. We also describe and evaluate the hybrid sentence alignment method we are using.

0 Introduction

There are only a dozen large languages with a hundred million speakers or more, accounting for about 40% of the world population, and there are over 5,000 small languages with less than half a million speakers, accounting for about 4% (Grimes 2003). In this paper we discuss some ideas about how to build parallel corpora for the five hundred or so medium density languages that lie between these two extremes based on our experience building a 50M word sentence-aligned Hungarian-English parallel corpus. Throughout the paper we illustrate our strategy mainly on Hungarian (14m speakers), also mentioning Romanian (26m speakers), and Slovenian (2m speakers), but we emphasize that the key factor leading the success of our method, a vigorous culture of native language use and (digital) literacy, is by no means restricted to Central European languages. Needless to say, the density of a language (the availability of digitally stored material) is predicted only imperfectly by the population of speakers: major Prakrit or Han dialects, with tens, sometimes hundreds, of million speakers, are low density, while minor populations, such as the Inuktitut, can attain high levels of digital literacy given the political will and a conscious Hansard-building effort (Martin et al 2003). With this caveat, population (or better, GDP) is a very good approximation for density, on a par with web size.

The rest of the paper is structured as follows. In Section 1 we describe our methods of corpus collection and preparation. Our hybrid sentence-level aligner is discussed in Section 2. Evaluation is the subject of Section 3.

1 Collecting and preparing the corpus

Starting with Resnik (1998), mining the web for parallel corpora has emerged as a major technique, and between English and another high density language, such as Chinese, the results are very encouraging (Chen and Nie 2000, Resnik and Smith 2003). However, when no highly bilingual domain (like `.hk` for Chinese or `.ca` for French) exists, or when the other language is much lower density, the actual number of automatically detectable parallel pages is considerably smaller: for example, Resnik and Smith find less than 2,000 English-Arabic parallel pages for a total of 2.3m words.

For medium density languages parallel web pages turn out to be a surprisingly minor source of parallel texts. Even in cases where the population and the economy is sizeable, and a significant monolingual corpus can be collected by crawling, mechanically detectable parallel or bilingual web pages exist only in surprisingly small numbers. For example a 1.5 billion word corpus of Hungarian (Halácsy et al 2004), with 3.5 million unique pages, yielded only 270,000 words (535 pages), and a 200m word corpus of Slovenian (202,000 pages) yielded only 13,000 words (42 pages) using URL parallelism as the primary matching criterion as in PTMiner (Chen and Nie 2000).

Some indication of this problem can already be seen in the low number, 2,491, of English–French pages by Resnik and Smith (2003), who discuss the issue under the heading "Too little data" (p 374). Since by GDP France is about 21 times the size of Hungary, and 66 times the size of Slovenia, we expect that an effort similar to ours would yield a quite respectable English-French parallel corpus, perhaps 5-6 m words for French, consistent with the growth of `.fr` since 1998. However, for medium density languages, even if we extrapolate optimistically for the next 5-10 years, the yield can not be expected to be significant.

Web pages are undoubtedly valuable for a diversity of styles and contents that is greater than what could be expected from any single source, but a few hundred web pages alone fall short of a sensible parallel corpus. Therefore, one needs to resort to other sources, many of them impossible to find by mechanical URL comparison, and often not even accessible without going through dedicated query interfaces. We discuss the nature of these resources using Hungarian as our primary example.

Literary texts The Hungarian National Library maintains a large public domain digital archive *Magyar Elektronikus Könyvtár* 'Hungarian Electronic Library' `mek.oszk.hu/indexeng.phtml` with many classical texts. Comparison with the Project Gutenberg archives at `www.gutenberg.org` yielded well over a hundred parallel texts by authors ranging from Jane Austen to Tolstoy. Equally importantly, many works still under copyright were provided by their publishers under the standard research exemption clause. While we can't publish most of these texts in either language, we publish the aligned sentence pairs alphabetically sorted. This "shuffling" somewhat limits usability inasmuch as higher than sentence-level text layout becomes inaccessible, but at the same time makes it prohibitively hard to reconstruct the original texts and contravene the copyright. Since shuffling nips copyright issues in the bud, it simplifies the complex task of disseminating aligned corpora considerably.

Religious texts The entire Bible has been translated to over 400 languages and dialects, and many religious texts from the Bhagavad Gita to the Book of Mormon enjoy nearly as broad currency. The Catholic Church makes a special effort to have papal edicts translated to other languages from the original Latin (see `www.vatican.va/archive`).[1]

[1] It has often been noted that archaic biblical texts offer little help in translating e.g. newswire text. The situation can be greatly improved by using a contemporary English translation (as opposed to the King James Version).

International Law From the Universal Declaration of Human Rights (`www.unhchr.ch/udhr`) to the Geneva Convention many important legal documents have been translated to hundreds of languages and dialects. Those working on the languages of the European Union have long availed themselves of the CELEX database.

Movie captioning Large mega-productions are often dubbed, but smaller releases will generally have only captioning, often available for research purposes. For cult movies there is also a vigorous subgenre of amateur translations by movie buffs.

Software internationalization Multilingual software documentation is increasingly becoming available, particularly for open source packages such as KDE, Gnome, OpenOffice, Mozilla, the GNU tools, etc (Tiedemann and Nygaard 2004).

Bilingual magazines Both frequent flyer magazines and national business magazines are often published with English articles in parallel. Many magazines from Scientific American to National Geographic have editions in other languages, and in many countries there exist magazines with complete mirror translations (for instance, *Diplomacy and Trade Magazine* publishes every article both in Hungarian and English).

Annual reports, corporate home pages Large companies will often publish their annual reports in English as well. These are usually more strictly parallel than the rest of their web pages.

There is no denying that the identification of such resources, negotiating for their release, downloading, format conversion, and character-set normalization remain labor-intensive steps, with good opportunities for automation only at the final stages. But such an effort leverages exactly the strengths of medium density languages: the existence of a joint cultural heritage both secular and religious, of national institutions dedicated to the preservation and fostering of culture, of multinational movements (particularly open source) and multinational corporations with a notable national presence, and of a rising tide of global business and cultural practices. Altogether, the effort pays off by yielding a corpus that is two-three orders of magnitude larger, and covering a much wider range of jargons, styles, and genres, than what could be expected from parallel web pages alone. Table 1 summarizes the different types of texts and their sizes in our Hungarian-English parallel corpus.

source	docs	E words (m)	H words (m)
Literary	156	14.6	11.5
Legal	10374	24.1	18.3
Captioning	437	2.5	1.9
Sw docs	187	0.8	0.7
Magazines	107	0.3	0.3
Business	19	0.5	0.4
Religious	122	2.3	2.0
Web	435	0.3	0.2
Total	**11550**	**44.6**	**34.6**

Table 1: Distribution of text types in the Hungarian–English parallel corpus

In addition to the texts, we identified other significant lexical resources, such as public domain glossaries specifically prepared for EU law, Microsoft software, Linux, and other particular domains and most importantly, a large (over 254,000 records) general-purpose bilingual dictionary manually created over many years by Attila Vonyó. Since there is no guarantee that such materials are available for other languages, in the next section we describe a sentence-alignment algorithm which does not rely on the existence of such bilingual dictionaries, but can take advantage of it if it is available.

After some elementary format-detection and conversion routines such as `catdoc` and `pdftotext` which are standard in the open source world, we have a corpus of raw text consisting of assumed parallel documents. While the texts themselves were collected and converted predominantly manually, the aligned bicorpus is derived by entirely automatic methods. Due to the manual effort, parallelism is nearly perfect, therefore the size of the raw corpus of collected texts is not significantly different from the size of the useful (aligned) data.

The first steps of our corpus preparation pipeline are tokenizers performing sentence and paragraph boundary detection and word tokenization. These are relatively simple flex programs (along the lines of Mikheev 2002) both for English and Hungarian. For languages with more complex morphology such as Hungarian, it makes sense to conflate by stemming morphological variants of a lexeme before the texts are passed to the aligner. We used `hunmorph`, a language-independent word analysis toolkit (Trón et al 2005) both for Hungarian and English.

The most important ingredient of the pipeline is of course automatic sentence alignment which we carried out using our own algorithm and software `hunalign`, described in detail in the next section.

2 Sentence level alignment

There are three main approaches to the problem of corpus alignment at the sentence level: length-based (Brown et al 1991, Gale and Church 1991), dictionary- or translation based (Chen 1993, Melamed 1996, Moore 2002), and partial similarity-based (Simard and Plamondon 1998). This last method in itself may work well for Indo-European languages (probably better between English and Romanian than English and Slovenian), but for Hungarian the lack of etymological relation suggests that the number of cognates will be low. Even where the cognate relationship is clear, as in *computer/kompjúter, strike/sztrájk* etc., the differences in orthography make it hard to gain traction by this method. Therefore, we chose to concentrate on the dictionary and length-based methods, and designed a hybrid algorithm, `hunalign,` that successfully amalgamates the two.

In the first step of the alignment algorithm, a crude translation of the source text is produced by converting each word token into the dictionary translation that has the highest frequency in the target corpus, or to itself in case of lookup failure.

This pseudo target language text is then compared against the actual target text on a sentence by sentence basis. The similarity score between a source and a target sentence consists of two major components: token-based and length-based. The dominant term of the token-based score is the number of shared words in the two sentences, normalized with the larger token count of the two sentences. A separate reward term is added if the proportion of shared numerical tokens is sufficiently high in the two sentences (especially useful for the alignment of legal texts).

For the length-based component, the character counts of the original texts are incremented by one, and the score is based on the ratio of longer to shorter. The relative weight of the two components was set so as to maximize precision on the Hungarian–English training corpus, but seems a sensible choice for other languages as well. Paragraph boundary markers are treated as sentences with special scoring: the similarity of two paragraph-boundaries is a high constant, the similarity of a paragraph-boundary to a real sentence is minus infinity, so as to make paragraph boundaries pair up.

The similarity score is calculated for every sentence pair around the diagonal of the alignment matrix (at least a 500-sentence neighborhood is calculated or all sentences closer than 10% of the longer text). This is justified by the observation that the beginning and the end of the texts are considered aligned and that the sentence ratio in the parallel text represents the average one-to-many assignment ratio of alignment segments, from which no significant deviations are expected. We find that 10% is high enough to produce reassuringly high recall figures even in the case of faulty parallelism such as long surplus chapters.

Once the similarity matrix is obtained for the relevant sentence pairs, the optimal alignment trail is selected by dynamic programming, going through the matrix with various penalties assigned to skipping and coalescing sentences. The score of skipping is a fixed parameter, learnt on our training corpus while the score of coalescing is the sum of the minimum of the two token-based scores and the length-based score of the concatenation of the two sentences. For performance reasons, the dynamic programming algorithm does not take into account the possibility of more than two sentences matching one sentence. After the optimal alignment path is found, a postprocessing step iteratively coalesces a neighboring pair of one-to-many and zero-to-one segments wherever the resulting new segment has a better character-length ratio than the starting one. With this method, any one-to-many segments can be discovered.

The hybrid algorithm presented above remains completely meaningful even in the total absence of a dictionary. In this case, the crude translation will be just the source language text, and sentence-level similarity falls back to surface identity of words.After this first phase a simple dictionary can be bootstrapped on the initial alignment. From this alignment, the second phase of the algorithm collects one-to-one alignments with a score above a fixed threshold. Based only on all one-to-one segments, cooccurrences of every source-target token pair are calculated. These, when normalized with the maximum of the two tokens' frequency yield an association measure. Word pairs with association higher than 0.5 but are are used as a dictionary.

Our algorithm is similar in spirit to that of Moore (2002) in that they both combine the length-based method with some kind of translation-based similarity. In what follows we discuss how Moore's algorithm differs from ours.

Moore's algorithm has three phases. First, an initial alignment is computed based only on sentence length similarity. Next, an IBM 'Model I' translation model (Brown et al 1993) is trained on a set of likely matching sentence pairs based on the first phase. Finally, similarity is calculated using this translation model, combined with sentence length similarity. The output alignment is calculated using this complex similarity score. Computation of similarity using Model I is rather slow, so only alignments close to the initially found alignment are considered, thus restricting the search space drastically.

Our simpler method using a dictionary-based crude translation model instead of a full IBM translation model has the very important advantage that it can exploit a bilingual lexicon, if one is available, and tune it according to frequencies in the target corpus or even enhance it with extra local dictionary bootstrapped from an initial phase. Moore's method offers no such way to tune a preexisting language model. This limitation is a real one when the corpus, unlike the news and Hansard corpora more familiar to those working on high density languages, is composed of very short and heterogeneous pieces. In such cases, as in web corpora, movie captions, or heterogeneous legal texts, average-based models are actually not close to any specific text, so Moore's workaround of building language models based on 10,000 sentence subcorpora has little traction.

On top of this, our translation similarity score is very fast to calculate, so the dictionary-based method can be used already in the first phase where a much bigger search space can be traversed. If the lexicon resource is good enough for the text, this first phase already gives excellent alignment results.

Maximizing alignment recall in the presence of noisy sentence segmentation is an important issue, particularly as language density generally correlates with the sophistication of NLP tools, and thus lower density implies poorer sentence boundary detection. From this perspective, the focus of Moore's algorithm on one-to-one alignments is less than optimal, since excluding one-to-many and many-to-many alignments may result in losing substantial amounts of aligned material if the two languages have different sentence structuring conventions.

While speed is often considered a mundane issue, `hunalign`, written in C++, is at least an order of magnitude faster than Moore's implementation (written in Perl), and the increase in speed can be leveraged in many ways during the building of a parallel corpus with tens of thousands of documents. First, rapid alignment allows for more efficient filtering of texts with low confidence alignments, which usually point to faulty parallelism such as mixed order of chapters (as we encountered in *Arabian Nights* and many other anthologies), missing appendices, extensive extra editorial headers (typical of Project Gutenberg), comments, different prefaces in the source texts etc. Once detected automatically, most cases of faulty parallelism can be repaired and the texts realigned. Second, debugging and fine-tuning lower-level text processing steps (such as the sentence segmentation and tokenization steps) may require several runs of alignment in order to monitor the impact of certain changes on the quality of alignment. This makes speed an important issue. Interestingly, runtime complexity of Moore's program seems to be very sensitive to the faults in parallelism. Adding a 300 word surplus preface to one side of *1984* but not the other slows down this program by a factor of five, while it has no detectable impact on `hunalign`.

Finally, Moore's aligner, while open source and clearly licensed for research, is not free software. In particular, parallel corpora aligned with it can not be made freely available for commercial purposes. Since we wanted to make sure that our corpus is available for any purpose, including commercial use, Moore's aligner program was not a viable choice.

3 Evaluation

In this section we describe our attempts to assess the quality of our parallel corpus by evaluating the performance of the sentence aligner on texts for which manually produced alignment is available. We also compare our algorithm to Moore's (2002) method.

Evaluation shows `hunalign` has very high performance: generally it aligns incorrectly at most a handful of sentences. As measured by Moore's method of counting only on one-to-one sentence-pairs, precision and recall figures in the high nineties are common. But these figures are overly optimistic because they hide one-to-many and many-to-many errors, which actually outnumber the one-to-one errors. In *1984,* for example, 285 of the 6732 English sentences or about 4.3% do not map on a unique Hungarian, and 716 or 10.6% do not map on a unique Romanian sentence – similar proportions are found in other alignments, both manual and automatic.

To take these errors into account, we used a slightly different figure of merit, defined as follows. The alignment trail of a text can be represented by a ladder, i.e. an array of pairs of sentence boundaries: rung (i, j) is present in the ladder iff the first i sentences on the left correspond to the first j sentences on the right. Precision and recall values are calculated by comparing the predicted and actual rungs of the ladder: we will refer to this as the *complete rung* count as opposed to the *one-to-one* count. In general, complete rung figures of merit tend to be lower than one-to-one figures of merit, since the task of getting them right is more ambitious: it is precisely around the one-to-many and many-to-one segments of the text that the alignment algorithms tend to stumble.

Table 3 presents precision and recall figures based on all the rungs of the entire ladder against the manual alignment of the Hungarian version of Orwell's 1984 (Dimitrova et al 1998).

condition	precision	recall
id	34.30	34.56
id+swr	74.57	75.24
len	97.58	97.55
len+id	97.65	97.42
len+id+swr	97.93	97.80
dic	97.30	97.08
len+dic-stem	98.86	98.88
len+dic	99.34	99.34
len+boot	99.12	99.18

Table 3: Performance of the sentence-level aligner under various conditions

If length-based scoring is switched off and we only run the first phase without a dictionary, the system reduces to a purely identity based method we denote by *id*. This will still often produce positive results since proper nouns and numerals will "translate" to themselves. With no other steps taken, on *1984 id* yields 34.30% precision at 34.56% recall. By the simple expedient of stopword removal, *swr*, the numbers improve dramatically, to 74.57% precision at 75.24% recall. This is due to the existence of short strings which happen to have very high frequency in both languages (the two predominant false cognates in the Hungarian-English case were *a* 'the' and *is* 'too').

Using the length-based heuristic *len* instead of the identity heuristic is better, yielding 97.58% precision at 97.55% recall. Combining this with the identity method does not yield significant improvement (97.65% precision at 97.55% recall). If, on top of this, we also perform stopword removal, both precision (97.93%) and recall (97.80) improve.

Given the availability of a large Hungarian-English dictionary by A. Vonyó, we also established a baseline for a version of the algorithm that makes use of this resource. Since the aligner does not deal with multiword tokens, entries such as *Nemzeti Bank* 'National Bank' are eliminated, reducing the dictionary to about 120k records. In order to harmonize the dictionary entries with the lemmas of the stemmer, the dictionary is also stemmed with the same tool as the texts. Using this dictionary (denoted by *dic* in the Table) without the length-based correction results in slightly worse performance than identity and length combined with stop word removal.

If the translation-method with the Vonyó dictionary is combined with the length-based method (*len+dic*), we obtain the highest scores 99.34% precision at 99.34% recall on rungs (99.41% precision and 99.40% recall on one-to-one sentence-pairs). In order to test the impact of stemming we let the algorithm run on the non-stemmed text with a non-stemmed dictionary (*len+dic-stem*). This established that stemming has indeed a substantial beneficial effect, although without it we still get better results than any of the non-hybrid cases.

Given that the dictionary-free length-based alignment is comparable to the one obtained with a large dictionary, it is natural to ask how the algorithm would perform with a bootstrapped dictionary as described in Section 2. With no initial dictionary but using this automatically bootstrapped dictionary in the second alignment pass, the algorithm yielded results (*len+boot*), which are, for all intents and purposes, just as good as the ones obtained from combining the length-based method with our large existing bilingual dictionary (*len+dic*). This is shown in the last two lines of Table 3. Since this method is so successful, we implemented it as a mode of operation of `hunalign`.

To summarize our results so far, the pure sentence length-based method does as well in the absence of a dictionary as the pure matching-based method does with a large dictionary. Combining the two is ideal, but this route is not available for the many medium density languages for which bilingual dictionaries are not freely avaliable. However, a core dictionary can automatically be created based on the dictionary-free alignment, and using this bootstrapped dictionary in combination with length-based alignment in the second pass is just as good as using a human-built dictionary for this purpose. In other words, the lack of a high-quality bilingual dictionary is no impediment to aligning the parallel corpus at the sentence level.

While we believe that an evaluation based on all the rungs of the ladder gives a more realistic measure of alignment performance, for the sake of correct comparison with Moore's method, we present some results using the one-to-one alignments metric. Table 4 summarizes results on Orwell's *1984* for Hungarian–English (*1984-HE-S*, stemmed and *1984-HE-U*, unstemmed), Romanian–English (*1984-RE-U*, unstemmed), as well as on Steinbeck's *Cup of Gold* for Hungarian–English (*CoG-HE-S*, 80k words, stemmed) using `hunalign` (with bootstrapped dictionary, no further tuning and omitting paragraph information) and Moore's (2002) algorithm (with the default values).

task	`hunalign`		`Moore '02`	
	prec	**rec**	**prec**	**rec**
1984-HE-S	99.22	99.24	99.42	98.56
1984-HE-U	98.88	99.05	99.24	97.39
1984-RE-U	97.10	97.98	97.55	96.14
CoG-HE-S	97.03	98.44	96.45	97.53

Table 4: Comparison of `hunalign` and Moore's (2002) algorithm on three texts. Performance figures are based on one-to-one alignments only.

In order to be able to compare the Hungarian and Romanian results for *1984*, we provide the Hungarian case for the unstemmed *1984*. One can see that both algorithms show a drop of performance. This makes it clear that the drop in quality from Hungarian–English to Romanian–English can not be attributed to the fact that we tuned our system on the Hungarian case. As mentioned earlier, the Romanian translation has 716 non-one-to-one segments compared to the Hungarian translation's 285. Given both algorithm's preference to globally diagonal and locally one-to-one alignments, this difference in one-to-one alignments is likely to render the Romanian–English alignment a harder task.

In order to sensibly compare our results with that of Moore's, paragraph information was not exploited. `huntoken`, the sentence tokenizer we use is able to identify paragraph boundaries which are then used by the aligner. Experiments showed that paragraph information can substantially improve alignment scores: measured on the Hungarian–English alignment of Steinbeck's 'Cup of Gold', the number of incorrect alignments drop from 148 to 115.[2] Therefore the figures shown in Table 4 are in no way absolute best bisentence scores for the texts in question.

4 Conclusion

In the past ten years, much has been written on bringing modern language technology to bear on low density languages. At the same time, the bulk of commercial research and product development, understandably, concentrated on high density languages. To a

[2] Although paragraph identification itself contains a lot of errors, improvement may be due to the fact that paragraphs, however faulty, are consistent in terms of alignment. The details of this and the question of exploiting higher-level layout information is left for future research.

surprising extent this left the medium density languages, spoken by over half of humanity, underresearched. In this paper we attempted to address this issue by proposing a methodology that does not shy away from manual labor as far as the data collection step is concerned. Harvesting web pages and automatically detecting parallels turns out to yield only a meager slice of the available data: in the case of Hungarian, less than 1%. Instead, we proposed several other sources of parallel texts based on our experience with creating a 50 million word Hungarian–English parallel corpus.

Once the data is collected and formatted manually, the subsequent steps can be almost entirely automated. Here we have demonstrated that our hybrid alignment technique is capable of efficiently generating very high quality sentence alignments with excellent recall figures, which helps to get the maximum out of small corpora. Even in the absence of any language resources, alignment quality is very high, but if stemmers or bilingual dictionaries are available, our aligner can take advantage of them.

Acknowledgements

The project is supported by an ITEM 2003 grant from the Hungarian Ministry of Informatics. We are grateful to Magyar Telecom for hardware and logistical support. We are also indebted to Tamás Váradi, and the whole Corpus Linguistics Department at the Institute of Linguistics, Hungarian Academy of Sciences, for the 1984 corpus and joint work on the Steinbeck text, and to Attila Vonyó for his Hungarian-English dictionary.

References

(Brown *et al.* 91) Peter Brown, Jennifer Lai, and Robert Mercer. Aligning sentences in parallel corpora. In *Proceedings of ACL29*, pages 169–176, 1991.

(Chen & Nie 00) Jiang Chen and Jian-Yun Nie. Automatic construction of parallel english-chinese corpus for cross-language information retrieval. In *Proceedings of the sixth conference on Applied natural language processing*, pages 21–28, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

(Chen 93) Stanley F. Chen. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st conference on Association for Computational Linguistics*, pages 9–16, Morristown, NJ, USA, 1993. Association for Computational Linguistics.

(Dimitrova *et al.* 98) Ludmila Dimitrova, Tomaz Erjavec, Nancy Ide, Heiki Jaan Kaalep, Vladimir Petkevic, and Dan Tufis. 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, 1998. Morgan Kaufmann Publishers.

(Gale & Church 91) William A. Gale and Kenneth Ward Church. A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184, 1991.

(Grimes 03) Barbara Grimes. *The Ethnologue (14th Edition)*. SIL, 2003.

(Halácsy *et al.* 04) Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Creating open language resources for Hungarian. In *Proceedings of Language Resources and Evaluation Conference (LREC04)*. European Language Resources Association, 2004.

(Martin *et al.* 03) Joel Martin, Howard Johnson, Benoit Farley, and Anna Maclachlan. Aligning and using an english-inuktitut parallel corpus. In *HLT-NAACL Workshop: Building and Using Parallel Texts*, pages 115–118, Edmonton, Alberta, Canada, May 31 2003. Association for Computational Linguistics.

(Melamed 00) I. Dan Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.

(Mikheev 00) Andrei Mikheev. Periods, capitalized words, etc. *Computational Linguistics*, 28(3):289–318, 2000.

(Moore 02) Robert C. Moore. Fast and accurate sentence alignment of bilingual corpora. In *Proc 5th AMTA Conf: Machine Translation: From Research to Real Users*, pages 135–244, Langhorne, PA, 2002. Springer.

(Resnik & Smith 03) Philip Resnik and Noah Smith. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380, 2003.

(Resnik 98) Philip Resnik. Parallel strands: A preliminary investigation into mining the web for bilingual text. In D. Farwell, L. Gerber, and E. Hovy, editors, *Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas*, Langhorne, PA, 1998. Springer.

(Simard *et al.* 98) Simard, Michel, and Pierre Plamondon. Bilingual sentence alignment: Balancing robustness and accuracy. In *Machine Translation*, volume Volume 13, no. 1, pages 59–80, 1998.

(Tiedemann & Nygaard 04) Jörg Tiedemann and Lars Nygaard. The opus corpus - parallel and free. In *Proceedings of LREC'04*, volume IV, pages 1183–1186, Lisbon, 2004.

(Trón *et al.* 05) Viktor Trón, György Gyepessi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. Hunmorph: open source word analysis. In *Proceeding of ACL*, 2005. paper presented at the ACL05 Software Workshop.