# Etudes in Chinese-Hungarian Corpus-Based Lexical Acquisition

Gábor Ugray

zydeodict@gmail.com

**Abstract:** The paper reports on a series of experiments to extract matching lexical items from a 6.1 million segment corpus of movie subtitles in Mandarin Chinese and Hungarian, with the aim of expanding an existing bilingual dictionary. The challenges of data cleansing and tokenization are outlined, and the outcome of word alignment, vector space embeddings, neural machine translation and two standard statistical approaches is presented. A bilingual concordance tool for end users, based on word alignments, is introduced. A quantitative and qualitative evaluation of the results finds that the new methods drastically outperform simple collocation extraction, but also shows that human judgement is indispensable before including vocabulary in a published dictionary.

## 1 Introduction

The last few years have brought two developments with promising consequences for digital lexicography. The first is the emergence of large bilingual corpora, even for an uncommon language pair such as Chinese-Hungarian. The second is neural-network-based machine learning driven by affordable GPUs. In this paper I report on a series of experiments to harness these developments for the expansion of CHDICT[1] [16], an open-source Chinese-Hungarian dictionary initially conceived as a translation of CC-CEDICT[2].

My work builds on OpenSubtitles2016 [9], a corpus of movie subtitles with 6.1 million Chinese-Hungarian segment pairs. I investigate the usefulness and limitations of word alignment, vector space models (VSM), neural machine translation (NMT), and statistical collocation extraction, to acquire lexical information from the corpus. All of these approaches have proven to be valuable sources of lexicographical insight, with VSM and NMT grossly outperforming simpler statistical methods. Furthermore, word alignment enables a bilingual concordance tool that is itself valuable for a broad audience.

[1] https://chdict.zydeo.net/en/

[2] https://www.mdbg.net/chinese/dictionary; https://cc-cedict.org/wiki/

## 2 Pre-processing

### 2.1 Data hygiene

The corpus that this work builds on contains 6.1 million aligned movie subtitles. I subsequently refer to the corpus's units as segments. These are usually, but not always, full sentences, which the corpus's authors aligned chiefly on the basis of timestamps. That method inevitably introduces some noise into the data through misaligned pairs, which is complemented by dirty data in the form of encoding errors, content in the wrong language, and for Chinese, in the wrong script (traditional instead of simplified).

I applied a mix of strategies to fix and prune the data. This included (a) converting to simplified if the segment included traditional-only characters, using OpenCC[3]; (b) fixing Hungarian *õ* and *û*; (c) discarding pairs where the text contained *{[]}@*, indicating escape sequences; (d) discarding pairs where the ratio of source and target length was beyond a threshold, allowing for greater variance in shorter segments; (e) discarding segments where the proportion of punctuation characters exceeded a threshold; (f) discarding pairs where too many Hungarian words were left unanalyzed by the emMorph morphological analyzer [11]; (g) discarding pairs where the Chinese contains characters that are neither in the Latin nor the Chinese script; and (h) removing duplicates.

After this preparation, the remaining corpus used throughout the exercises contains 2.9 million segment pairs.

### 2.2 Chinese word segmentation[4]

A key challenge for any Chinese NLP task is the lack of word delimiters in written Mandarin. There is no single universally accepted word segmentation method, and as we will see, the optimal approach depends on the task at hand – including, even, treating each character as a separate token.

I am aware of two available segmenters, ICTCLAS[5] [18] and Jieba.[6] The first was used by Brysbaert et al. to obtain their corpus-based word frequencies for SUBTLEX-CH [2]. Unfortunately I was unable to compile and execute this tool, but in some places I rely on its output indirectly through the published SUBTLEX-CH word frequencies. Both ICTCLAS and Jieba are hybrid tools combining a dictionary and Hidden Markov Models.

Additionally, I experimented with an algorithm inspired by Gensim's [12] *Phraser*. Starting with individual characters, it iteratively merges adjoining units that co-occur more frequently than predicted by chance. My purpose was to prevent a perceived

[3] https://github.com/BYVoid/OpenCC by BYVoid
[4] Word segmentation is not to be confused with segments, the corpus's sentence-level units.
[5] https://github.com/NLPIR-team/NLPIR/tree/master/NLPIR%20SDK/NLPIR-ICTCLAS
[6] https://github.com/fxsjy/jieba by Sun Junyi (2013)

over-eagerness of Jieba in joining measure words to determiners; complements to verbs; and compound nouns. Somewhat predictably, this home-grown approach underperformed Jieba, and I abandoned the experiment. I refer to this method as "Exp-merge" later in the text.

Since the goal is to acquire lexical units that humans expect to find in a dictionary, an established large dictionary's headword list is a good benchmark. For a sense of how a segmenter's "idea" of words corresponds to the dictionary's judgement, see *Figure 1* below.

To obtain this chart, I first created three word lists, ranked by frequency as measured on a corpus segmented by the three tools. The SUBTLEX-CH frequencies are those published by Brysbaert et al. from a 33 million word corpus. The others are my own calculations on the pruned bilingual corpus. The figure is a histogram, showing values for 100 points on the X axis, each representing a 1000-word bucket in the ranked list of the 100,000 most frequent words. The Y values indicate the dictionary coverage of each bucket: how many of those 1000 words are found in CC-CEDICT.

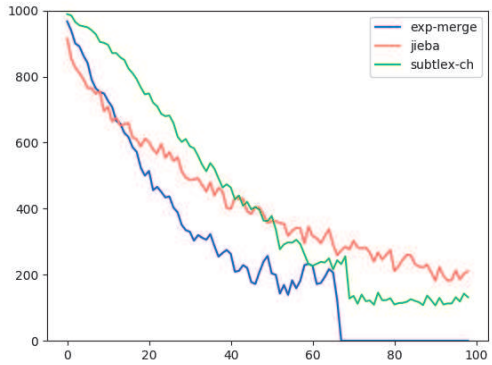


**Figure 1:** Lexical coverage of CC-CEDICT's 107k headwords, depending on the choice of tokenizer.

For comparison, Exp-merge produced 66k distinct words; Jieba's output contains 238k distinct words; SUBTLEX-CH's list has 100k words. Conversely, of CC-CEDICT's 107k simplified headwords, only 27.2k, 54.6k and 42.6k are attested in the Exp-Merge, Jieba and SUBTLEX-CH frequency lists, respectively. The diagram and these figures show that all segmenters disagree greatly both among themselves and with CC-CEDICT about the "definition" of words in Mandarin Chinese.

### 2.3 Hungarian stemming and tokenization

The case for stemming Hungarian to avoid the data scarcity problem is evident. I used the HFST-based [8] emMorph [11] analyzer and processed its output with my own C#

port of the *Stemmer* class and its dependencies from the GATE wrapper[7] [13]. I used a naïve disambiguation rule, always opting for the shortest stem. If the shortest stem was a particle+verb or a compound, I subsequently treated those as separate tokens.

Unfortunately HFST proved to be prohibitively slow in practice; I had to analyze the corpus's 772k surface forms from the command line and use an in-memory dictionary. Stemming produced 63.8k distinct stems, of which 20,237 occur at least 3 times. 162k surface forms, of 20.9% of the total, were left unanalyzed. These are overwhelmingly typos or non-Hungarian proper names.

For certain tasks I used byte-pair encoding (BPE) [14] as an alternative tokenization method. BPE gained popularity in recent years in NMT systems because it addresses the closed-vocabulary problem, albeit at the cost of being an arbitrary and not linguistically motivated approach.

## 3 Word-aligned bilingual concordance

Word alignment of the training corpus has been a staple of statistical MT from the outset [1], and it gained relevance again as an aid to the attention mechanism in NMT [3]. The alignment method itself has been improved significantly as recently as 2013 by the authors of *fast_align* [5]. The approach has great appeal because it can combine corpus-wide co-occurrence probabilities with local sequence information in individual segment pairs.

The temptation to build a Linguee-like[8] tool for searching the cleansed bilingual corpus, word-aligned on a segment level, was irresistible. My initial aim was to create a research tool for dictionary authoring, but as we will see in *Results*, the outcome has immediate value for end users too.

### 3.1 Training

I executed *fast_align* after tokenizing both Chinese and Hungarian in several different ways. *Table 1* shows *fast_align*'s reported perplexity values for these combinations.

For Hungarian tokenization, *bpe20k*, *bpe30k* and *bpe40* refer to BPE with 20k, 30k and 40k merges, respectively. *surf-lo* stands for no stemming, only lower-case normalization. *stem-lo* stands for lower-cased stems. For Chinese, Jieba outperforms the experimental word segmenter. Interestingly, increasing BPE's output vocabulary leads to worse outcomes.

BPE with Jieba allows for a few insightful searches, where a Chinese preposition or verbal complement is (correctly) mapped to a Hungarian suffix. However, *stem-lo+Jieba* grossly outperforms all other combinations, and was chosen for the final tool.

[7] https://github.com/dlt-rilmta/hunlp-GATE/tree/master/Lang_Hungarian/resources/hfst/hfst-wrapper

[8] https://www.linguee.de/

| Chinese word segmenter | Hungarian tokenizer | Perplexity |
|---|---|---|
| Jieba | bpe20k | 284.48 |
| Jieba | bpe30k | 295.85 |
| Jieba | bpe40k | 300.16 |
| Jieba | surf-lo | 216.47 |
| Jieba | stem-lo | 86.97 |
| Exp-merge | bpe20k | 355.75 |
| Exp-merge | stem-lo | 112.44 |

**Table 1:** *fast_align*'s reported perplexity values after training for 5 iterations, depending on the choice of tokenizer.

### 3.2 Presentation

I integrated a custom-developed tool for searching the word-aligned bilingual corpus within the CHDICT website[9]. For an illustration, see *Figure 2*, with a few results for 汽车 qìchē. The tool allows searching for either Chinese or Hungarian text, and presents matching segment pairs from the corpus.


**Figure 2:** A few sample search results from the word-aligned bilingual concordance tool's output.

In the results, the search term and its matches in the opposite language are highlighted in each segment. I used a slightly modified version of *fast_align* that outputs confidence values, which are indicated by the strength of the highlight. If the Chinese search term happens to be a substring of a Jieba token, the full token is also shown with a lighter highlight to clarify what the alignment truly means.

For Hungarian searches the tool uses two separate indexes. One matches the query's exact surface form; the other matches stems. Because of HFST's performance issues, queries are "stemmed" through an auxiliary table mapping the corpus's 772k surface forms to their chosen stems. The tool uses Sphinx[10] for quick and memory-efficient indexing and retrieval.

[9] https://chdict.zydeo.net/en/corpus
[10] http://sphinxsearch.com/

## 4 Bilingual word embeddings

Vector Space Models (VSMs) [15] embody the idea that a word's paradigmatic and semantic properties can be captured by quantifying what other words they tend to co-occur with. TF-IDF has been widely used in document retrieval since the 1970s paper that Gerard Salton never wrote [4]. More recently, neural networks have been used to learn word embeddings [10], replacing the closed formulas based on term counts.

I attempted to extract translations by embedding words from two languages in a single vector space. The approach is similar to Vulić et al. [17], but while that work relies on sampling non-sentence-aligned document pairs, my corpus allowed creating bags of words directly from Chinese+Hungarian segment pairs.

### 4.1 Extraction

The standard way to build a term-context matrix is to observe a small window up to about a dozen words in monolingual text. For my experiment I created, instead, a single bag of words from each Chinese and Hungarian segment pair. For clarity I prefixed Hungarian tokens with *hu_* and Chinese ones with *zh_*, although the languages can easily be distinguished by script.

To create the word embeddings, I used Gensim's Word2Vec model in skip-gram mode, with a window beyond the largest combined segment length. Gensim is an efficient re-implementation of Mikolov et al.'s neural word2vec model [10]. I used 200 dimensions, a value lower than the 300-500 that is standard in neural MT systems.

To define what a "word" is, I used Jieba for Chinese and lower-cased word stems for Hungarian. In this task, I discarded BPE because it is not linguistically informed.

Once the word vectors were learned, I applied a brute-force quadratic search to find the 40 nearest (by cosine similarity) Hungarian words with a frequency of 3 or higher, for each of the 54.6k CC-CEDICT headwords that are attested in the Jieba-segmented data.

### 4.2 Filtering

The raw output was, predictably, extremely noisy. Scores of 0.8 or higher are very reliable, but only 2,673 Chinese words have such a close Hungarian neighbor. On the other hand, spurious Hungarian matches tend to recur often in the top 40 list of several Chinese words. These proliferous matches are invariably noise: *nore* tops the list, showing up with 7,060 Chinese headwords, followed by *tada*, *csatlakozatok*, *lndítsuk* and *áilítólag*.

This enabled a filtering approach that also keeps potentially useful matches with a lower score. After ignoring Hungarian words that occur in the top 40 list of at least 100 different Chinese words, I was left with shorter non-empty lists for 34k CC-CEDICT headwords.

### 4.3 Outcome

Vector similarity delivers on its promise, returning a collage of words related in various ways. The list is always mixed: apart from the remaining noise, it contains semantic equivalents; complementary parts of frequent collocations; or simply vaguely related concepts. For illustration, here is the list of 遗物 yíwù, which CC-CEDICT glosses as *remnant:*

0.58 holmi • 0.53 ereklye • 0.50 mamaji • 0.50 hamvaszt • 0.49 felipe • 0.49 gyűjtemény • 0.48 régiség • 0.48 taiáitam • 0.48 drágakő • 0.48 hagyaték • 0.47 davenport • 0.47 yukio • 0.46 szuvenír • 0.46 ékszer • 0.46 mohammad • 0.46 amun-ra • 0.45 josemaría • 0.45 itthagyott • 0.44 coggins • 0.44 anyakönyvi • 0.44 bizsu • 0.43 hamu • 0.43 irat • 0.43 régész • 0.43 tárgy

This impressionistic collage helps disambiguate *remnant* into the eventual Hungarian glosses: *maradvány; ereklye; tárgyi emlék; hagyaték*. The vector space is a weak and noisy source of "translations" as such, but it has proven very valuable as a lexicographical tool to chart a headword's associations, connotations and register.

## 5 Neural MT

Google MT was already one of several sources for the compilation of CHDICT's original 11k entries [16]. Direct Chinese-Hungarian translations were rarely useful, with strong hints that Google uses English as a pivot language. I now investigated if custom NMT models trained from a Chinese-Hungarian corpus would yield useful headword translations.

I used OpenNMT [7] to train several models on data tokenized in different ways. All models have word embeddings of 500 dimensions and a 500-node 2-layer RNN. They were trained for 13 epochs with SGD, an initial learning rate of 1, annealed at a factor of 0.7 starting at epoch 9. Each model took approximately 12 hours to train on an NVIDIA GTX 1080 GPU, with mini-batches of 256. *Table 2* shows the perplexity and BLEU score results.

| Seg-ZH | Tok-HU | Perplexity | BLEU |
|---|---|---|---|
| **chars** | **bpe10k** | **14.10** | **9.58** |
| chars | bpe20k | 16.27 | 8.85 |
| chars | bpe40k | 18.35 | 8.80 |
| **chars** | **stem-lo** | **22.46** | **10.41** |
| exp-merge | bpe20k | 16.13 | 9.99 |
| **Jieba** | **bpe20k** | **16.29** | **10.03** |

**Table 2**: Final perplexity values and BLUE scores reported by OpenNMT, depending on the choice of segmenter/tokenizer.

If the goal were machine translation of full sentences, these results would be underwhelming. My aim, however, was only to extract helpful Hungarian hints for individual Chinese words. To this end I translated CC-CEDICT's 107k simplified headwords with the three models highlighted above, using a beam size of 10 and keeping the 20 best results.

### 5.1 Outcome

The output frequently shows anomalies that are well known to NMT practitioners. One example is the "I don't know problem"[11] also observed in neural chatbots, where the system defaults to a generally likely target segment. From the film subtitles corpus, this produces output like *igen; mi; igen uram; nem; szia;* etc. The other salient anomaly occurs with BPE-segmented Hungarian, where the system gets stuck in repetition loops: *ho hoho; hohohoho; hohohohoho;* etc.

Because the prediction score produced by inference is not a good indicator of quality, I used the same filtering approach as with vector similarities, discarding target strings that recur for many inputs. In fact I applied a stricter filter and discarded all of an engine's results for a given input if the first (best-scored) translation was on the proliferous noise list.

For illustration, this is the filtered output of the 3 selected engines for the previously mentioned headword, 遗物 yíwù:

**MT char-char**
örökre • marad • maradvány • maradt • maradj • maradsz • örökség • maradványokat • hagyja • holmija • egy maradvány • maradványok • öröksége
**MT char-stem**
tárgy • örökség • ereklye • hagyaték • maradvány • egy ereklye • kincs • rom • zsákmány • holmi • egy tárgy • egy vagyon • búcsú • tulajdon • sajnál • vagyon
**MT jie-char**
tárgy • cucc • tárgyról • holmi • tárgya • egy tárgy • tárgyak • dolgokat • tárgyakat • dolgok • tárgyat • holmik • dolgot • ez a tárgy

This is a drastic improvement over Google MT, which translates 遗物 as *emlékei*. In the cases where a Chinese headword has direct equivalents in Hungarian, the most frequent ones usually show up among the translations of multiple engines.

A major advantage over the other approaches is that MT is occasionally capable of producing translations consisting of multiple tokens, such as compound words or short expressions.

[11] forum.opennmt.net/t/english-chatbot-advice/32/5

## 6 Collocation classics

Collocation extraction using established statistical formulas has little novelty value, but I also included it in my effort as a source of easily obtainable candidates. I gleaned Chinese-Hungarian token pairs with two scoring methods, log-likelihood and mutual information. In this case I used Jieba and lower-cased stems for tokenization, these being the only linguistically motivated ones.

Just as before, I relied on a frequency threshold and excessive proliferation for filtering, instead of an arbitrary score threshold. After filtering, LL and MI produced non-empty candidate lists for 16.7k and 31.8k CC-CEDICT headwords, respectively.

## 7 Results

Two factors make a quantitative evaluation of the investigated methods complicated. First, the very aim of these exercises is to aid in the expansion of a pioneering bilingual dictionary, which means that there is no *a priori* ground truth available.

To work around this fact, I selected a batch of entries from CC-CEDICT, picking headwords that had filtered candidates from VSM; from at least 2 MT engines; and at least 1 collocation method. There were 9k headwords matching these criteria, of which I randomly sampled 400. I then proceeded to manually compile their Chinese-Hungarian entries, consulting CC-CEDICT's English glosses as well as the candidates from the new corpus-based extraction methods, the concordance tool and other sources. This created a *post-hoc* ground truth to benchmark against.

The second complication results from the fact that a dictionary entry, even in CHDICT's simplistic format, is not a flat list of target-language equivalents. Entries are structured into senses, which in turn may contain multiple alternatives, plus meta-information in parentheses. *Figure 3* illustrates this. Evaluating flat candidate lists against a structured gold standard is not straightforward.

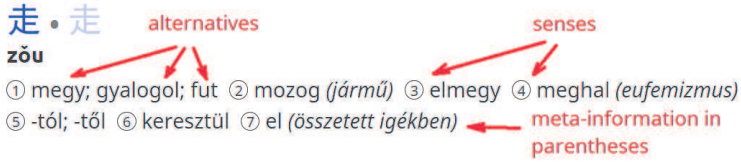


**Figure 3:** A sample entry from CHDICT, as it appears to end users in the live website.

### 7.1 Quantitative evaluation

Because of the complications outlined above, I had to resort to custom definitions in order to measure recall and precision. This makes the figures somewhat difficult to

compare to other efforts. They are, however, quite useful for comparing the different approaches within this paper's context. *Table 3* presents the results.

| Method | R-XS1 | R-XSX | R-ASX | P-1SX |
|---|---|---|---|---|
| Bilingual VSM | 31% | 40% | 21% | 31% |
| MT-char-char | 28% | 50% | 28% | 28% |
| MT-char-stem | 34% | 55% | 35% | 34% |
| MT-jie-char | 32% | 58% | 31% | 32% |
| Colloc/log-likelihood | 11% | 12% | 4% | 11% |
| Colloc/mutual-information | 14% | 25% | 10% | 14% |

**Table 3**: Recall and precision of the investigated vocabulary extraction methods. Recall measures: R-XS1: At least one alternative in the manually compiled entry is at the top of the candidate list. R-XSX: At least one alternative appears somewhere on the list. R-ASX: All alternatives appear somewhere on the list. Precision measure: The top candidate appears among the alternatives in the manually compiled entry.

The table's numbers are based on 390 manually prepared CHDICT entries; 10 Chinese tokens were rejected as dictionary headwords altogether. The retained entries contain a total of 576 senses, 784 alternatives, and 156 parenthesized remarks or labels.

VSM grossly outperforms the two conventional collocation extraction methods. The three NMT engines appear to have different strengths and weaknesses, depending on the metric, but they significantly outperform even VSM as a source of actual translations.

### 7.2 Benefits and limitations in the lexicographical process

*Figure 4* shows how candidates are presented in the lexicographical workbench during the compilation of entries. All the Hungarian words from the lists on the right are also added to an auto-complete dictionary to speed up typing.

I did not include word alignments among the hints; instead, the concordance tool itself can be invoked with a shortcut.

The workbench logs the time spent compiling each headword. It is beyond this paper's scope to analyze these logs, but it appears the enriched information does not affect the speed of lexicographical work. It contributes greatly, instead, to the confidence and breadth of the Hungarian glosses produced.


**Figure 4:** Information shown in the lexicographical workbench for the translation of a single Chinese headword.

Often, but not always, the candidates from the different methods include the words eventually selected for the Hungarian glosses, or otherwise help explore the Chinese headword's uses and meanings. A human can normally identify the remaining noise on the lists, and the relevant items condense what would otherwise be the result of hours of corpus discovery and hunting for attestations.

They candidates are not, however, reliable enough to be included in the dictionary without human judgement. Using only candidates with a very high score would leave an unacceptably small number of reliable matches, and miss important but less frequent senses. Lowering the threshold, in turn, would result in excessive noise or a proliferation of candidates. Finally, unsupervised methods obviously fail when a Chinese lexical item can only be paraphrased and when the target equivalent needs disambiguating remarks.

### 7.3 Augmented dictionary

The word-aligned bilingual search tool has proven to be the most versatile approach. Its value in the lexicographical process is evident, as it allows researching real-life contexts in which a headword has been attested.

But why should such research be limited to lexicographical work? The dictionary's end users benefit equally from a chance to browse headwords in context, discovering autonomously a word's translations along with typical collocations and associations.

When integrated in a dictionary, the search tool is a substitute for example phrases, which are particularly labor-intensive to compile.

Additionally, as a fallback when a word is not found in the dictionary, the tool enables end users to discover its meaning from the translated sentence pairs. The coverage of dictionaries is limited by the person-years needed to compile them,

especially for rare combinations like Chinese-Hungarian. A large bilingual corpus inevitably encodes more knowledge than is humanly possible to compile.

## References


1. Brown, Peter F., Cocke, John, Della Pietra, Stephen A., Della Pietra, Vincent J., Jelinek, Fredrick, Lafferty, John D., Mercer, Robert L., and Roossin, Paul S.: A Statistical Approach To Machine Translation. In Computational Linguistics, Volume 16 Issue 2, June 1990, pp. 79-85.
2. Cai, Q., Brysbaert, M.: SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. PLoS ONE 5(6): e10729. doi:10.1371/journal.pone.0010729 (2010)
3. Crego, Josep et al.: SYSTRAN's Pure Neural Machine Translation Systems. arXiv:1610.05540
4. Dubin, David: The Most Influential Paper Gerard Salton Never Wrote. In: Library Trends 52(4) Spring 2004: 748-764.
5. Dyer, C., Chahuneau, V., and Smith, N. A.: A simple, fast, and effective reparameterization of IBM model 2. In Proceedings of the 2013 Conference of the North American Chapter of the ACM: Human Language Technologies, pp. 644-648, Atlanta, GA, USA (2013)
6. Halpern, J. and Kerman, J.: The Pitfalls and Complexities of Chinese to Chinese Conversion. In: *Proceedings of the 14th International Unicode Conference*, Cambridge, MA, March 1999
7. Klein, Guillaume; Kim, Yoon; Deng, Yuntian; Crego, Josep; Senellart, Jean; Rush, Alexander M.: OpenNMT: Open-source Toolkit for Neural Machine Translation. In Proceedings of ACL 2017
8. Lindén, K., Silfverberg, M., Pirinen, T.: HFST tools for morphology – an efficient open-source package for construction of morphological analyzers. In: Mahlow, C., Piotrowski, M. (eds.) State of the Art in Computational Morphology, Communications in Computer and Information Science, vol. 41, pp. 28–47. Springer Berlin. Heidelberg (2009)
9. Lison, Pierre and Tiedemann, Jörg: OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. NIPS 2013
11. Attila Novák; Borbála Siklósi; Csaba Oravecz: A New Integrated Open-source Morphological Analyzer for Hungarian. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, pp. 1315–1322.
12. Řehůřek, Radim and Sojka, Petr: Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*. Valletta, Malta: University of Malta, 2010. pp. 46-50.
13. Sass Bálint, Miháltz Márton, Kundráth Péter: Az e-magyar rendszer GATE környezetbe integrált magyar szövegfeldolgozó eszközlánca. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017) (2017)
14. Sennrich, Rico, Haddow, Barry and Birch, Alexandra: Neural Machine Translation of Rare Words with Subword Units Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany.
15. Turney, Peter D., Patrick, Pantel: From Frequency to Meaning: Vector Space Models of Semantics. In: Journal of Artificial Intelligence Research 37 (2010) pp. 141-188.

16. Ugray Gábor: Egy vakmerő digitális lexikográfiai kísérlet: a CHDICT nyílt kínai-magyar szótár. Poster presentation. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017) (2017)
17. Vulić, Ivan, Moens, Marie-Francine: Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction. In: Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP (Vol. 2), July 2015, pp. 719-715.
18. Zhang Hua-Ping, Yu Hong-Kui, Xiong De-Yi Xiong and Liu Qun: HHMM-based Chinese Lexical Analyzer ICTCLAS, proceedings of 2nd SigHan Workshop, July 2003, pp.184-187.