Automated Building of Sentence-Level Parallel Corpus and Chinese Hungarian Dictionary

Yidi Zhang CS 13' zyd413@gmail.com
Zhongxiu Liu CS 14' zhongxiuliu@wpi.edu

Advisors: Gabor Sarkozy, WPI
Andras Kornai, MTA-Sztaki

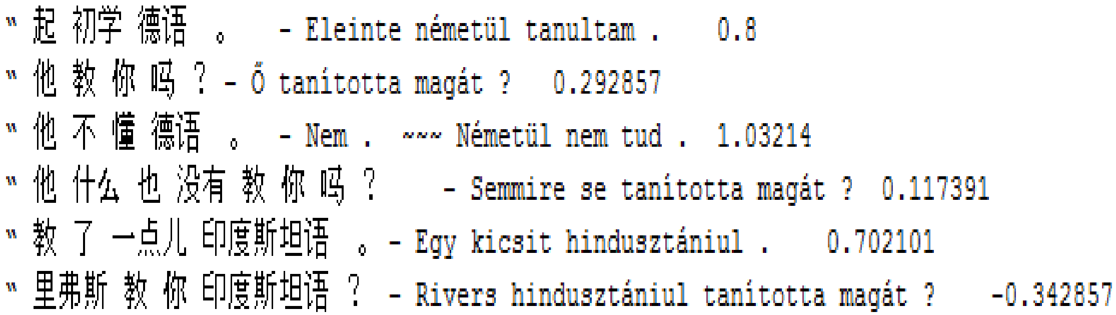
April 23rd, 2013

Outline

- Background
- Methodology
- Results & Evaluation
- Future Work

Background

- Parallel Corpus: Parallel corpus is text placed alongside its translation

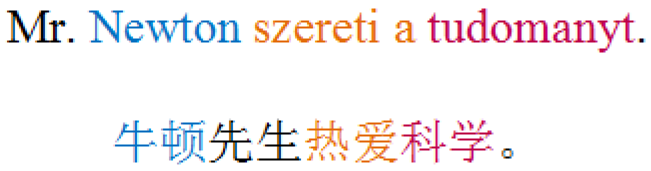


```
" 起 初学 德语 。   - Eleinte németül tanultam .    0.8
" 他 教 你 吗 ？ - Ő tanította magát ?   0.292857
" 他 不 懂 德语  。  - Nem .  ~~~ Németül nem tud .  1.03214
" 他 什么 也 没有 教 你 吗 ？    - Semmire se tanította magát ?  0.117391
" 教 了 一点儿 印度斯坦语  。- Egy kicsit hindusztániul .    0.702101
" 里弗斯 教 你 印度斯坦语 ？  - Rivers hindusztániul tanította magát ?    -0.342857
```

- Automatically created Dictionary
  - Automatically created dictionaries are in great demand
  - SZTAKI online Dictionary [1]

Background

- Difference between Hungarian and Chinese



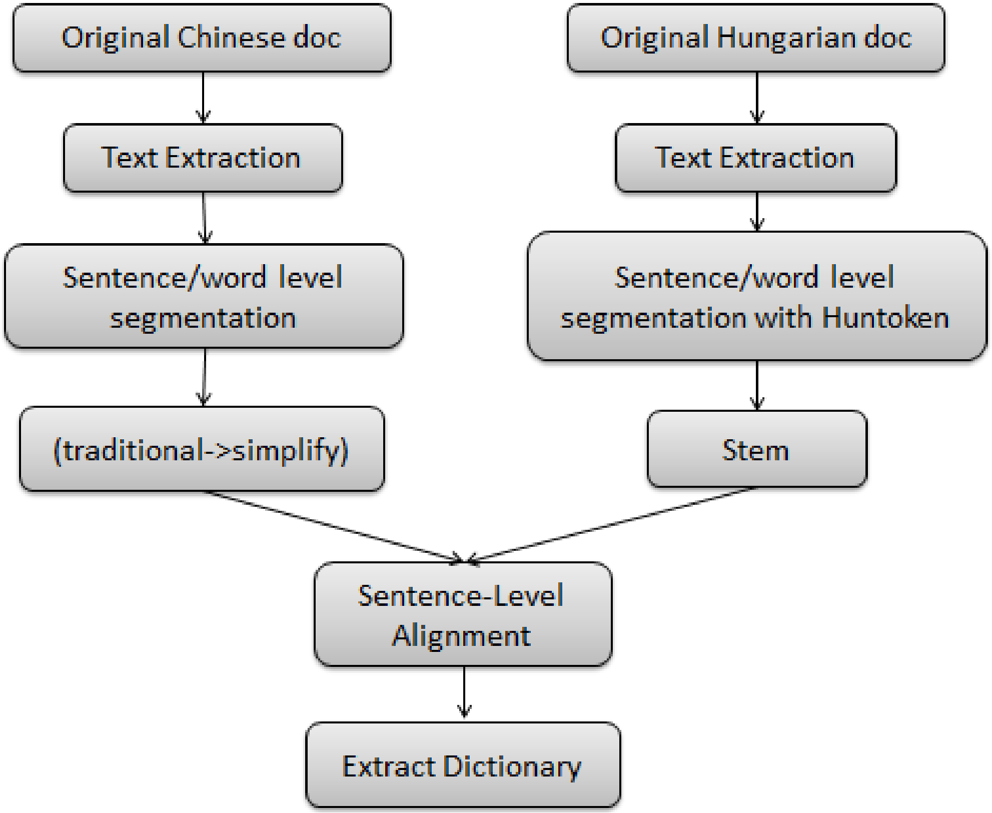
Mr. Newton szereti a tudomanyt.

牛顿先生热爱科学。

  - Medium-density vs. High-density
  - Word-based vs. Character-based
  - Other linguistic Differences
    - Chinese does not have time-tense, plural or other form change of characters
    - Chinese does not have space between words
    - Chinese has less ambiguous sentence-ending characters

Methodology

- Flowchart of our Methodology



Step1: Collecting Parallel Documents

- Variety
  - Literature, Subtitles, Religious texts
- Raw Documents "Clean-up"
  - Unifying file Encoding to UTF-8
    - Encoding_transform.py
  - Filtering out incomplete documents and useless
  - contents such as formatting tags in webpages
- Total: 60 Parallel Documents

Step2: Normalizing Hungarian Documents

- Sentence/Word Level Segmentation
  - Rule-based segmenter: Huntoken [2]

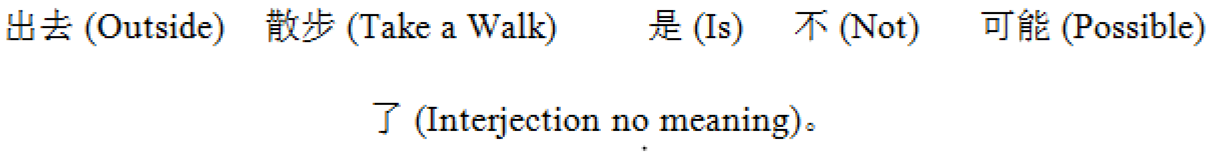
- Stemming
  - Extract root word in Hungarian
  - Stemming tool: Hunmorph[3]

babonás	babonás||ADJ	K	S
lett	van||VERB<PAST>	K	S

Step3: Normalizing Chinese Documents

- Sentence Level Segmentation
  - Rule-based segmenter: chinese-sentencizor.py

- Word Level Segmentation
  - Stanford Segmentation[4]
    - Input: 出去散步是不可能的了。
    - Output:



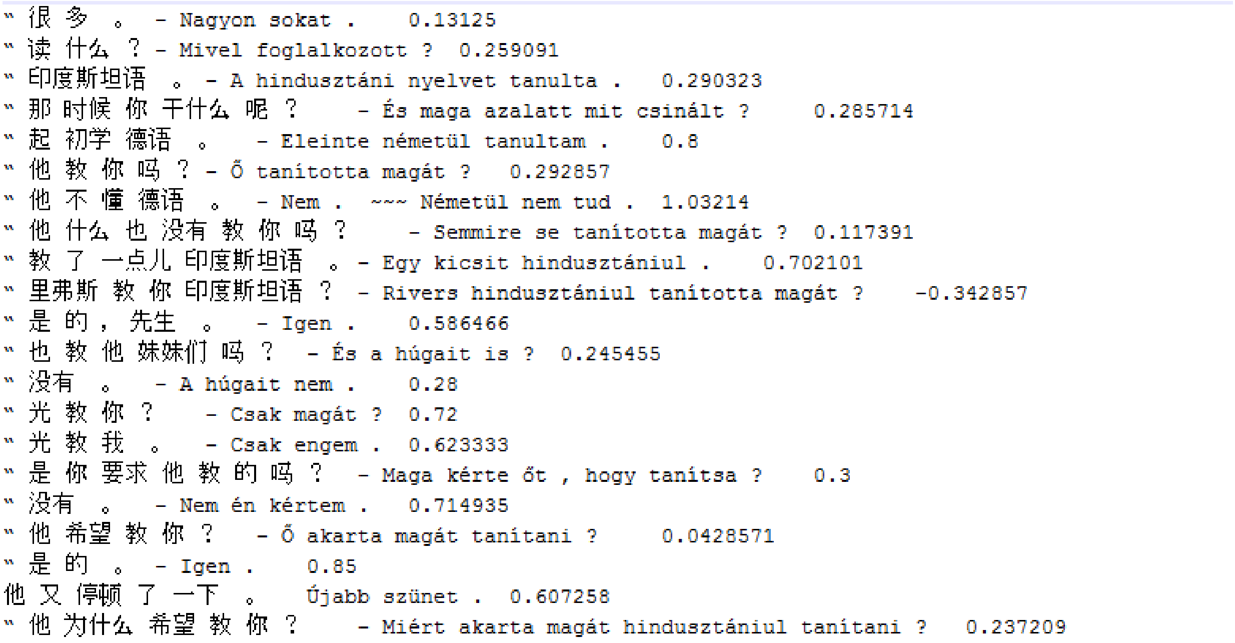
出去 (Outside) 散步 (Take a Walk) 是 (Is) 不 (Not) 可能 (Possible) 了 (Interjection no meaning)。

- Traditional Chinese to Simplified Chinese

Step4: Create Sentence-Level Parallel Corpus – Hunalign [5]

- Hunalign
  - Input: Parallel Documents

    Output:



```
“ 很 多  。 - Nagyon sokat .    0.13125
“ 读 什么 ? - Mivel foglalkozott ?  0.259091
“ 印度斯坦语  。 - A hindusztáni nyelvet tanulta .   0.290323
“ 那 时候 你 干什么 呢 ?     - És maga azalatt mit csinált ?     0.285714
“ 起 初学 德语  。    - Eleinte németül tanultam .    0.8
“ 他 教 你 吗 ? - Ő tanította magát ?   0.292857
“ 他 不 懂 德语  。   - Nem .  ~~~ Németül nem tud .  1.03214
“ 他 什么 也 没有 教 你 吗 ?     - Semmire se tanította magát ?  0.117391
“ 教 了 一点儿 印度斯坦语  。 - Egy kicsit hindusztániul .    0.702101
“ 里弗斯 教 你 印度斯坦语 ? - Rivers hindusztániul tanította magát ?    -0.342857
“ 是 的 , 先生  。   - Igen .    0.586466
“ 也 教 他 妹妹们 吗 ?  - És a húgait is ?  0.245455
“ 没有  。    - A húgait nem .    0.28
“ 光 教 你 ?    - Csak magát ?  0.72
“ 光 教 我  。    - Csak engem .  0.623333
“ 是 你 要求 他 教 的 吗 ?  - Maga kérte őt , hogy tanítsa ?    0.3
“ 没有  。    - Nem én kértem .   0.714935
“ 他 希望 教 你 ?   - Ő akarta magát tanítani ?     0.0428571
“ 是 的  。  - Igen .    0.85
他 又 停顿 了 一下  。    Újabb szünet .  0.607258
“ 他 为什么 希望 教 你 ?     - Miért akarta magát hindusztániul tanítani ?   0.237209
```

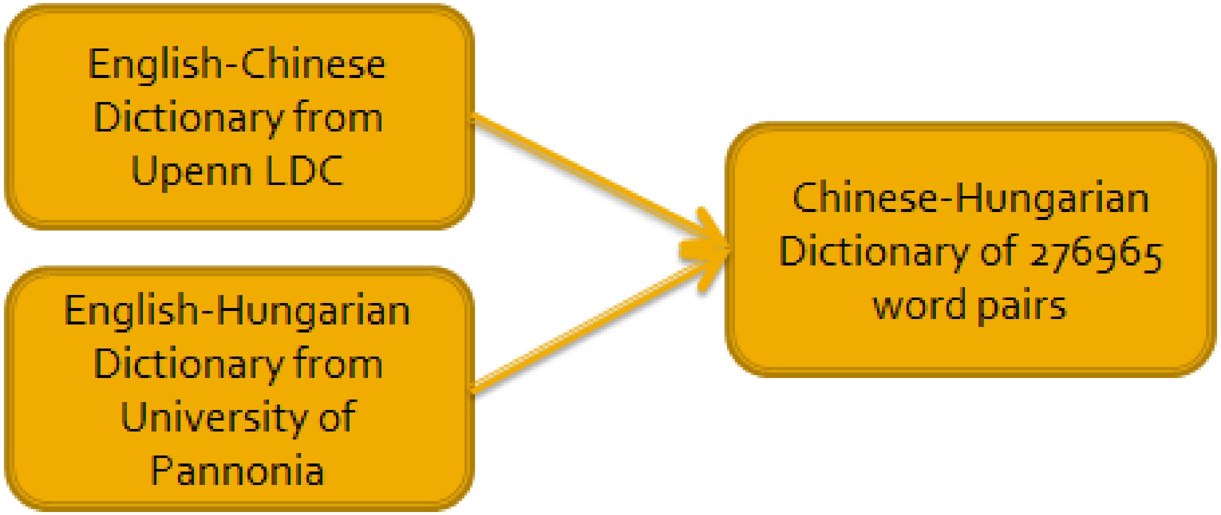
Step4: Create Sentence-Level Parallel Corpus – Hunalign [5]

- How does Hunalign work?
  - Calculate similarity score of each sentence pair
    - Token-based score = $\frac{num\ shared\ words}{larger\ token\ count\ of\ two\ sentences} + award$ *(for high proportion of shared tokens)*
    - Length-based score = $\frac{length(longer)+1}{Length(shorter)+1}$
  - Obtain similarity matrices, Calculate the best alignment
  - Post-processing: iteratively coalesces near sentence-pairs – detect one-to-many, many-to-one alignments.

Step4: Create Sentence-Level Parallel Corpus – Hunalign [5]

- Hunalign can gain help from bilingual dictionary

- Input dictionary we used:



English-Chinese Dictionary from Upenn LDC

English-Hungarian Dictionary from University of Pannonia

Chinese-Hungarian Dictionary of 276965 word pairs

Step4: Create Sentence-Level Parallel Corpus

- Run Hunalign on the 60 normalized parallel documents with upenn_upann_hu-zh.dict
- Run on Stemmed Documents -> Resulting in line-number pairs -> Parallel with unstemmed documents

Step5: Create Hungarian-Chinese Dictionary

- Hundict [6]
  - Dice score

$$QS = \frac{2|A \cap B|}{|A| + |B|}$$

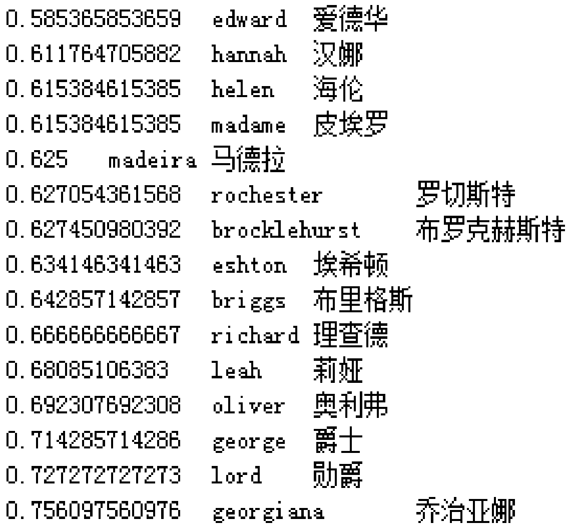
$$= \frac{2 * Co - Occurence}{(Zh - Occurence + Co - Occurence) + (En - Occurent + Co - Occurence)}$$

  - Final Options we used:
    - dice = 0.2
    - iter=5

Step5: Create Hungarian-Chinese Dictionary

- Run Hundict on the unstemmed sentence-level parallel corpus output by Hunalign

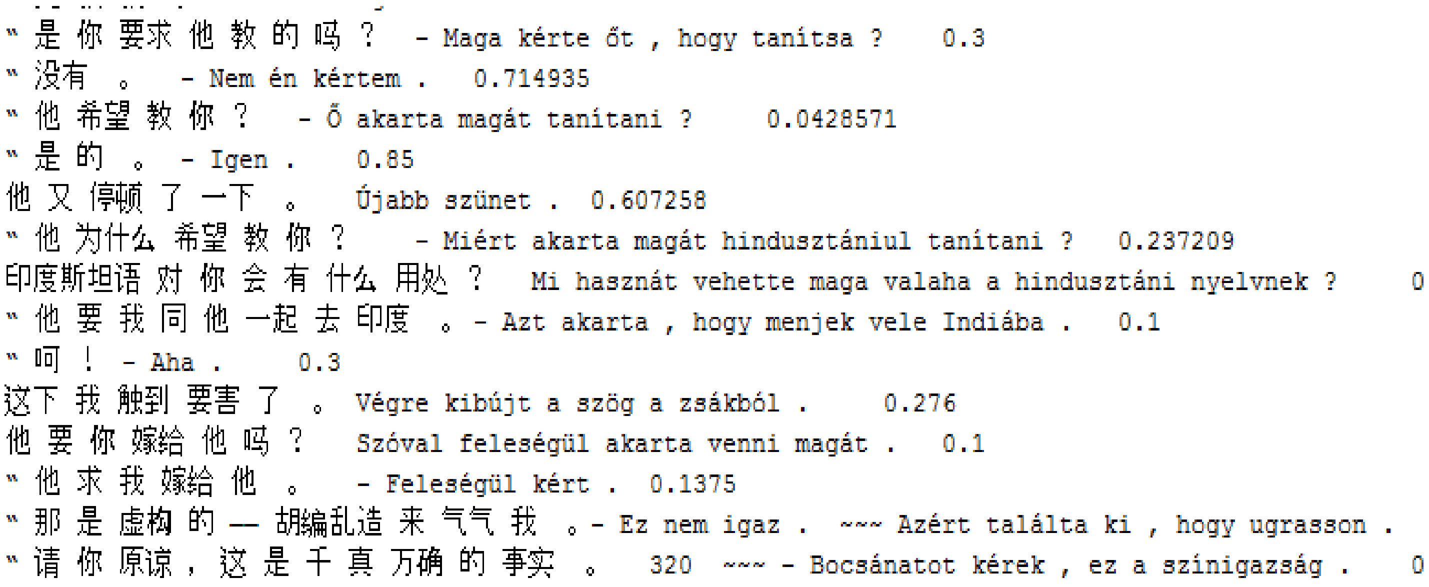
- Filtering out name pairs



```
0.585365853659	edward	爱德华
0.611764705882	hannah	汉娜
0.615384615385	helen	海伦
0.615384615385	madame	皮埃罗
0.625	madeira	马德拉
0.627054361568	rochester	罗切斯特
0.627450980392	brocklehurst	布罗克赫斯特
0.634146341463	eshton	埃希顿
0.642857142857	briggs	布里格斯
0.666666666667	richard	理查德
0.68085106383	leah	莉娅
0.692307692308	oliver	奥利弗
0.714285714286	george	爵士
0.727272727273	lord	勋爵
0.756097560976	georgiana	乔治亚娜
```

Result I: Sentence-level Parallel Corpus

- Output
  - Size: 26427 sentence pairs
  - Sample:



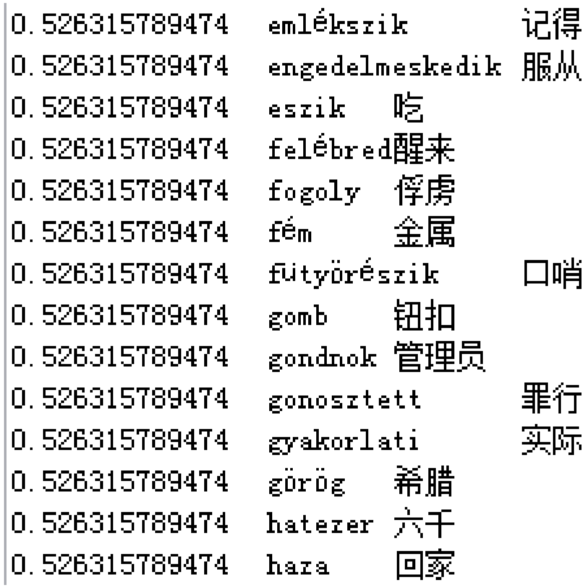
```
“ 是 你 要求 他 教 的 吗 ？   - Maga kérte őt , hogy tanítsa ?    0.3
“ 没有  。   - Nem én kértem .   0.714935
“ 他 希望 教 你 ？   - Ő akarta magát tanítani ?     0.0428571
“ 是 的  。  - Igen .    0.85
他 又 停顿 了 一下  。   Újabb szünet .  0.607258
“ 他 为什么 希望 教 你 ？     - Miért akarta magát hindusztániul tanítani ?   0.237209
印度斯坦语 对 你 会 有 什么 用处 ？   Mi hasznát vehette maga valaha a hindusztáni nyelvnek ?     0
“ 他 要 我 同 他 一起 去 印度  。 - Azt akarta , hogy menjek vele Indiába .   0.1
“ 呵 ！  - Aha .      0.3
这下 我 触到 要害 了  。  Végre kibújt a szög a zsákból .     0.276
他 要 你 嫁给 他 吗 ？   Szóval feleségül akarta venni magát .   0.1
“ 他 求 我 嫁给 他  。    - Feleségül kért .  0.1375
“ 那 是 虚构 的 — 胡编乱造 来 气气 我  。- Ez nem igaz .  ~~~ Azért találta ki , hogy ugrasson .
“ 请 你 原谅 , 这 是 千 真 万确 的 事实  。   320  ~~~ - Bocsánatot kérek , ez a színigazság .    0
```

Result I: Sentence-Level Parallel Corpus

- Evaluation
  - Score output by Hunalign is meaningless
  - We did manual checking:
    - Randomly select sample sentence pairs
    - Hungarian and Chinese native speakers communicate in English

Results II: Hungarian-Chinese Dictionary

- Output
  - Size: 23932 word pairs + 613 name pairs
  - Example:



```
0.526315789474	emlékszik	記得
0.526315789474	engedelmeskedik	服从
0.526315789474	eszik	吃
0.526315789474	felébred醒来
0.526315789474	fogoly	俘虏
0.526315789474	fém	金属
0.526315789474	fütyörészik	口哨
0.526315789474	gomb	钮扣
0.526315789474	gondnok	管理员
0.526315789474	gonosztett	罪行
0.526315789474	gyakorlati	实际
0.526315789474	görög	希腊
0.526315789474	hatezer	六千
0.526315789474	haza	回家
```

Results II: Hungarian-Chinese Dictionary

- Evaluation: Manual Checking

Similarity Score Range	% Word Pairs	Accuracy
[0, 0.1)	33.2%	7/20 ≈ 35%
[0.1,0.2)	19%	12/20 ≈ 60%
(0.2-0.4]	20.5%	17/20≈ 85%
(0.4-0.6]	16.1%	18/20≈ 90%
(0.6-0.8]	8.6%	20/20≈ 100%
(0.8-1.0]	2.6%	19/20≈ 95%

Results II: Hungarian-Chinese Dictionary

- Common Errors:

  - Translation Error

  draga: 亲爱 (dear, darling)   Right: 昂贵(dear, expensive)

  - Incomplete Translation

  Sellő： 鱼 (fish)   Right: 美人鱼 (mermaid)

  - Word Phrase

  Gyalázatos： 行为 (act)   Right:可耻(shameful)

Future Work and Improvements

- Introducing new dictionaries with our methodology
  - E.g. Hungarian-Japanese, Romanian-Chinese
- Collecting larger and better parallel documents
- Developing quantified evaluation for parallel corpora and bilingual dictionaries

User Guide

- Create you own sentence-level parallel corpus and dictionary
  - Step1: Download our package
  - Step2: runHunalign.sh
  - Step3: runHundict.sh

Acknowledgement

- **András Kornai,** *MTA-SZTAKI*
- **Gábor Sárközy,** *Worcester Polytechnic Institute*
- **Attila Zséder,** *MTA-SZTAKI*
- **Judit Ács,** *MTA-SZTAKI*
- **Katalin Pajkossy,** *MTA-SZTAKI*

Reference

[1] SZTAKI. <http://szotar.sztaki.hu/>

[2] http://mokk.bme.hu/resources/huntoken/

[3] Tron, V., Gyepesi, G., Halacsy, P., Kornai, A., Nemeth, L., Varga, D. Hunmorph: Open Source Word Analysis. *In Proceedings of the Association of Computational Linguistics.* Web. 2005. <ftp://ftp.mokk.bme.hu/LDC/doc/acl05software.pdf>

[4] Tseng, P., Chang, G., Andrew, G., Jurafsky, D., Manning, C. A Conditional Random Field Word Segmenter. *SIGHAN of the Association for Computational Linguistics.* Web. 2005. <http://nlp.stanford.edu/pubs/sighan2005.pdf>

[5] Varga, D., Halacsy, P., Kornai, A., Nagy, V., Nemeth, L., Tron, V. Parallel Corpora for Medium Density Languages. *Proceedings of Recent Advances in Natural Language Processing.* Web. 2005. <http://www.kornai.com/Papers/ranlp05parallel.pdf>.

[6] Zséder, A. SZTAKI. https://github.com/zseder/hundict

Thank You!