The Role of Parallel Corpora in Bilingual Lexicography

Enikő Héja
Research Institute for Linguistics, HAS
eheja@nytud.hu
Outline

• The project
• The role of parallel corpora in lexicography
• Workflow
• Results
• Conclusions and future work
EFNILEX (EFNIL)

• Objectives:
  • Dictionaries for human use covering every day vocabulary for medium density languages
  • 20,000-45,000 entries (depending on the size of available resources)

• Methodology:
  • Statistical word alignment
  • Based on parallel corpora

• Language pairs:
  • Hungarian – Slovenian
  • Hungarian – Lithuanian
Advantages

- Parallel corpus $\Rightarrow$ *Corpus-driven technique* to diminish the role of lexicographers’ intuition

- Usage-based, representative translations

- Clear ranking between more likely and less likely translations
  - Most-used translation equivalents are ranked higher (Example I)

- Provided contexts facilitate the creation of encoding dictionaries (Example II)

- Compilation of the reversed dictionary is more simple
Advantages – a Sample

- Positive evidence that the various sub-senses of a word are translated in the same way

<table>
<thead>
<tr>
<th>HUN LEMMA</th>
<th>LIT LEMMA</th>
<th>TRANSLATIONAL PROBABILITY</th>
<th>FREQUENCY OF HUN LEMMA</th>
<th>FREQUENCY OF LIT LEMMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Születik</td>
<td>Gimti (-sta,-é)</td>
<td>0.579005</td>
<td>169</td>
<td>174</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HUN</th>
<th>LIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ö 1870-ben született</td>
<td>Jis gimë 1870 metais</td>
</tr>
<tr>
<td>He was born in 1870</td>
<td></td>
</tr>
<tr>
<td>De Fache mintha erre született volna</td>
<td>Bet Fasas, regis, tiesiog tam gimës</td>
</tr>
<tr>
<td>As if Fache was born to do this</td>
<td></td>
</tr>
<tr>
<td>Hungarian</td>
<td>English</td>
</tr>
<tr>
<td>--------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Úgy látszik, szerencsétlen csillagzat alatt születél</td>
<td>It seems that you were born under an unlucky star</td>
</tr>
<tr>
<td>... mert ikrei születtek.</td>
<td>..., nes jai gimé dvynukai.</td>
</tr>
<tr>
<td>... because twins were born to her.</td>
<td></td>
</tr>
<tr>
<td>Maga úriembernek született.</td>
<td>Tu gimei dżentlemanu.</td>
</tr>
<tr>
<td></td>
<td>You was born a gentleman.</td>
</tr>
<tr>
<td>... hogy Buddha nem lótusvirágból született?</td>
<td>..., kad Buda gimé ne iš lotoso žiedo?</td>
</tr>
<tr>
<td></td>
<td>...that Buddha was born from a lotus flower?</td>
</tr>
</tbody>
</table>
Difficulties

• Creation of the parallel corpus is tedious

• Dictionaries generated by word alignment comprise only one-to-one mappings between lemmata
  
  • Does not handle MWEs, collocations, verbal constructions => can be added based on the provided contexts manually afterwards
Resources and Tools

• **Resources:** goal: a 10,000,000-token corpus for each language

• **Tools:** language dependent tools are needed for each language
  - Sentence splitting
  - Tokenising
  - Lemmatising
  - Disambiguating between lemmata
Resources

- Lithuanian-Hungarian, Slovenian-Hungarian
- Collecting direct translations yielded only moderate success
- Instead, translations from a third language
  - Parallel web pages from the web (~200,000 tokens per language).
  - Literature from the web (mainly resources of Hungarian digital archives: MEK, DIA)
  - Texts from national corpora
    - Lithuanian: Lithuanian National Corpus, Lithuanian-English parallel corpus
    - Slovenian: FIDA corpus
Tools

- Language specific tools were available in the form of tool-chains
  - LIT: Centre of Computational Linguistics, Vytautas Magnus University
  - SLO: Jozef Stefan Institute, freely available at http://nl.ijs.si/jos/analyse/
  - HUN: Research Institute for Linguistics, used for the annotation of the Hungarian National Corpus
Workflow

1. Evaluation
- "Raw corpus" for L1
  - Encoding issues
  - Line breaks
  - Normalization
- "Raw corpus" for L2
  - Encoding issues
  - Line breaks
  - Normalization

2. Evaluation
- Annotated corpus-L1
  - Tokenization
  - Sentence segmentation
  - Lemmatization
- Annotated corpus-L2
  - Tokenization
  - Sentence segmentation
  - Lemmatization

3. Evaluation
- Parallel corpus
- Sentence alignment

4. Evaluation
- Word alignment
- Dictionary for L1, L2

Creating dictionary by using word alignment and example sentences from corpora

Performed by language-independent tools

Performed by language-specific tools

Hunalign sentence aligner
(language independent)

GIZA++ wordaligner
(language independent)
The quality of the resulting dictionary depends highly on the factors below:

- Quality of input texts
- Quality of sentence alignment
- Quality of word alignment
## Size of Parallel Corpora

- **Lithuanian-Hungarian**

<table>
<thead>
<tr>
<th>Language</th>
<th>Tokens</th>
<th>Aligned Units (AU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lithuanian</td>
<td>1,765,000</td>
<td>147,158</td>
</tr>
<tr>
<td>Hungarian</td>
<td>2,121,000</td>
<td>147,158</td>
</tr>
</tbody>
</table>

- **Slovenian-Hungarian**

<table>
<thead>
<tr>
<th>Language</th>
<th>Tokens</th>
<th>Aligned Units (AU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slovenian</td>
<td>733,000</td>
<td>38,574</td>
</tr>
<tr>
<td>Hungarian</td>
<td>666,000</td>
<td>38,574</td>
</tr>
</tbody>
</table>
After word alignment we had the following data at our disposal:

| HUN LEMMA | LIT LEMMA | Translational probability $P(W_{target}|W_{source})$ | Corpus frequency HUN LEMMA | Corpus frequency LIT LEMMA |
|-----------|-----------|-----------------------------------------------------|-----------------------------|-----------------------------|
| Ajak (lip)| Lüpa      | 0.77063                                             | 312                         | 509                         |
| Alagút (tunel) | Tunelis | 0.755043                                             | 145                         | 157                         |

**Objective**: to find the “ideal” values for these parameters
Most Probable Translation Candidates II

- We set these values based on the evaluation of the HUN-SLO translation candidates

- Every lemma should occur at least 5 times => to have sufficient amount of data to give a reliable estimation of $P(tr)$

- If $P(tr) < 0.5$, the proportion of correct translation candidates drops considerably

- 65% of the translation candidates is correct
## Preliminary Results

<table>
<thead>
<tr>
<th></th>
<th>Number of Translation-Candidates Above the Threshold</th>
<th>Expected Number of Correct Translation-Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUNGARIAN-SLOVANIAN</td>
<td>4969</td>
<td>3230</td>
</tr>
<tr>
<td>HUNGARIAN-LITHUANIAN</td>
<td>4025</td>
<td>2616</td>
</tr>
</tbody>
</table>
Evaluation: Useful Translation Candidates

• Correct translational equivalents
  \[\text{gyümölcs} – \text{vaisius} \text{ (fruit)}\]

• Partially correct translational equivalents $\Rightarrow$ Post editing is needed
  - Improper lemmatization
  - Only partial match in the case of MWEs
    compounds \[\text{fofelügyelő} – \text{vyriausiasis inspektorius} \text{ (chief inspector)}\]
    collocations \[\text{bíborosi testület} – \text{Kardinolų kolegija} \text{ (cardinal college)}\]

• Looser semantic relation (e.g. hypernymy)
  \[\text{lúdtoll} \text{ (literally: goose-feather)} – \text{plunksna} \text{ (literally: feather, pen)}\]
  intended meaning in both cases: \emph{quill pen}
Evaluation: Useless Translation Candidates

- Irrelevant vocabulary (e.g. recurrent proper names) \([\text{Abdul} – \text{Abdulas}]\)

- Incorrect translation candidates
  - Usually due to the loose translations of texts
Evaluation – Data

- Out of 4025 HUN-LIT translation pair 863 pairs were sampled
  - freq ≥ 5, \( P(w_{\text{target}} \mid w_{\text{source}}) \geq 0.5 \)

- Evaluation intervals:
  - \( 0.5 \leq P(w_{\text{target}} \mid w_{\text{source}}) < 0.7 \)
  - \( 0.7 \leq P(w_{\text{target}} \mid w_{\text{source}}) < 1 \)
  - \( P(w_{\text{target}} \mid w_{\text{source}}) = 1 \)
### Results

<table>
<thead>
<tr>
<th></th>
<th>Useful candidates</th>
<th>Useless candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P(tr)</strong></td>
<td>OK</td>
<td>Post-editing</td>
</tr>
<tr>
<td>[0.5, 0.7]</td>
<td>52.1 %</td>
<td>32.9 %</td>
</tr>
<tr>
<td>Sum</td>
<td>Σ 85 %</td>
<td></td>
</tr>
<tr>
<td>[0.7, 1)</td>
<td>65.3 %</td>
<td>31.9 %</td>
</tr>
<tr>
<td>Sum</td>
<td>Σ 97, 2 %</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>38 %</td>
<td>13 %</td>
</tr>
<tr>
<td>Sum</td>
<td>Σ 51%</td>
<td></td>
</tr>
</tbody>
</table>

- Proportion of incorrect translation pairs is low
- **85 %** of translation pairs are *useful* in the 1. probability range
- **97.2 %** of translation pairs are *useful* in the 2. range
- P(tr) = 1 produces the lowest proportion of useful candidates and the highest ratio of irrelevant pairs
Related Meanings I

- *Presupposition*: frequent words tend to have more meanings than less frequent ones

- Lithuanian-Hungarian dictionary:
  - Frequency of Lithuanian lemma is min. 100
  - Translational probability was considerably decreased (0.5 $\Rightarrow$ 0.02)
## Related Meanings – Example I

| LIT       | HUN    | P(wi|wi) | ENG          |
|-----------|--------|--------|--------------|
| puikus    | jó     | 0.128  | good         |
| puikus    | remek  | 0.071  | great, all right |
| puikus    | tökéletes | 0.052 | perfect     |
| puikus    | szép   | 0.048  | nice         |
| puikus    | pompás | 0.035  | splendid     |
| puikus    | jól    | 0.035  | well         |
| puikus    | nagyszerű | 0.035 | great       |
| puikus    | finom  | 0.028  | fine         |
| puikus    | gyönyörű | 0.02   | marvelous    |

- **Puiku**, - atsakė balsas.  
- **Puikus** darbas.  
- **Remek** – válaszolta a hang. ( **All right** – the voice answered )  
- **Szép** munka volt. ( **Good** job )
Use in the creation of encoding dictionaries

- Tisztán, világosan, jól can modify verbs of perception with the same meaning
- Láthatóan refers to the fact that the emotional change a person underwent did not remain hidden
- Világosan is used with verbs of cognition and communication meaning that the content of the act is comprehensible
- Tisztán would mean that the speech conveying the message was clearly pronounced
Conclusion and Future Work

- The corpus-driven nature of this method decreases the role of human intuition during dictionary building

- Translations are provided together with their contexts

- Translations can be ranked according to their likelihood

- Size of parallel corpora has to be augmented

- Automatic treatment of MWEs, collocations and verbal constructions should be included in the workflow
Thank you for your attention!