Automatically Generated Online Dictionaries

Tamás Váradi and Enikő Héja

Research Institute for Linguistics
Hungarian Academy of Sciences

varadi, eheja@nytud.hu

Background I

- **Goals**:
  - Dictionaries for human use covering every day vocabulary for medium density languages
  - 20.000-45.000 entries (depending on the size of available resources)
- **Realization:**
  - According to the state-of-the-art there are no LT-methods that could fully eliminate lexicographic expertize during dictionary building
  - *Objective*: to provide lexicographers with automatically generated resources facilitating their work => Proto-dictionaries
  - Manual post-editing is needed!

Background II

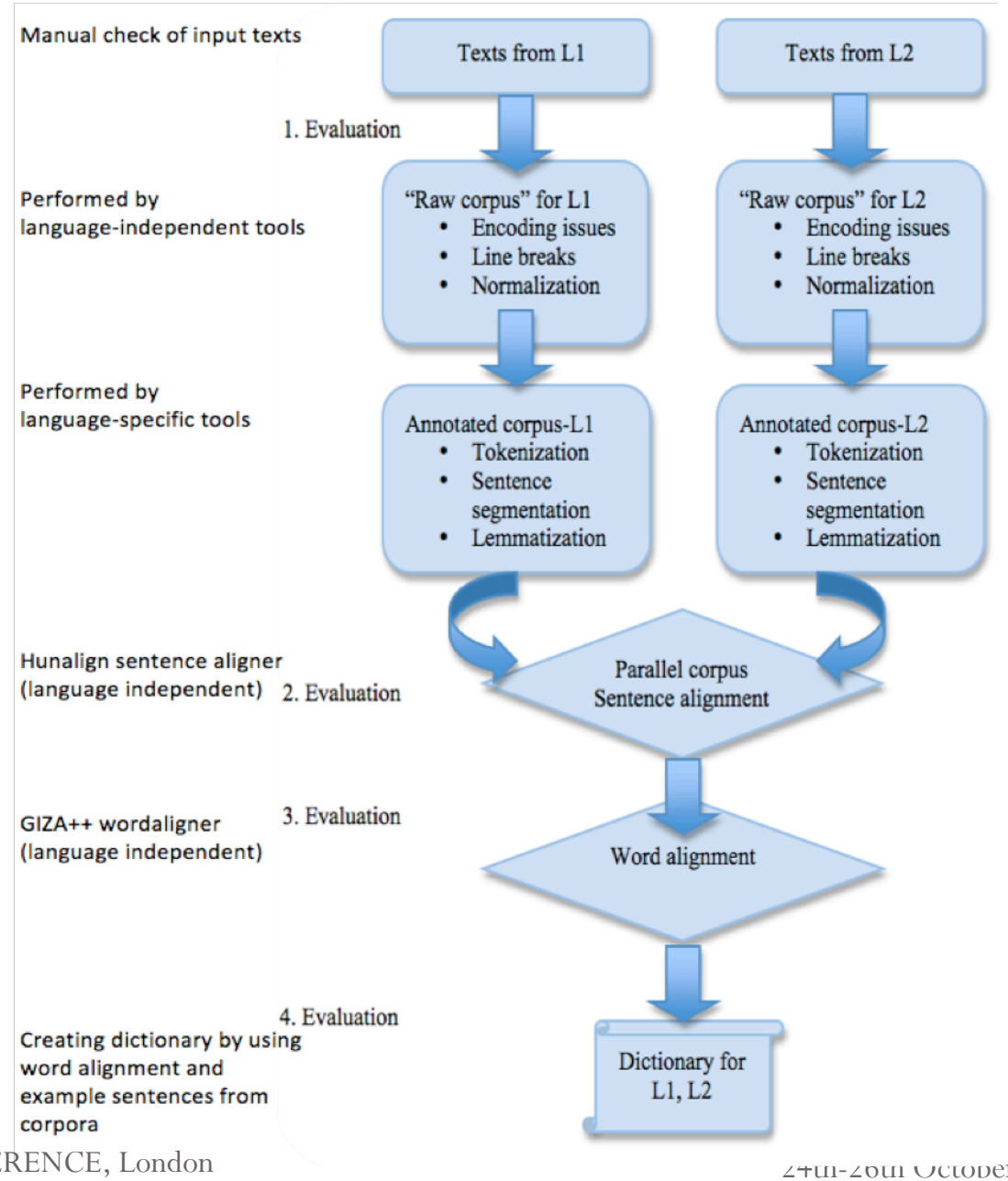
- **Methodology:**
  - Statistical word alignment
  - Based on sentence aligned parallel corpora
- **Language pairs**:
  - (Hungarian – Slovenian)
  - Hungarian – Lithuanian
  - French – Dutch

Proto-dictionary: an Example

HUN LEMMA	LIT LEMMA	TRANSLATIONAL PROBABILITY	FREQUENCY OF HUN LEMMA	FREQUENCY OF LIT LEMMA
Születik	Gimti (-sta,-ė)	0.579005	169	174

HUN	LIT
Ő **1870-ben született**	Jis **gimė 1870 metais**
He **was born in 1870**	
De Fache mintha **erre született** volna	Bet Fasas, regis, tiesiog **tam gimęs**
As if Fache **was born to** do this	

Workflow


Manual check of input texts
Texts from L1
Texts from L2
1. Evaluation
Performed by language-independent tools
"Raw corpus" for L1
Encoding issues
Line breaks
Normalization
"Raw corpus" for L2
Encoding issues
Line breaks
Normalization
Performed by language-specific tools
Annotated corpus-L1
Tokenization
Sentence segmentation
Lemmatization
Annotated corpus-L2
Tokenization
Sentence segmentation
Lemmatization
Hunalign sentence aligner (language independent)
2. Evaluation
Parallel corpus
Sentence alignment
GIZA++ wordaligner (language independent)
3. Evaluation
Word alignment
4. Evaluation
Creating dictionary by using word alignment and example sentences from corpora
Dictionary for L1, L2

Advantages

- provides consistent and reliable method
  - selecting source language headwords
  - finding the translational equivalents
    - Usage-based, representative translations
    - Clear ranking between more likely and less likely translations
    - Most-used translation equivalents are ranked higher
- Example sentences facilitate the creation of encoding dictionaries
- Reversing the dictionary is a more straightforward process

Recent Activities

- Size of proto-dictionaries has been increased
  - Bigger parallel corpus
  - Evaluation on the basis of refined parameters
- Proto-dictionary for French and Dutch
- Online dictionary query system to disseminate our results
- A proof-of-concept experiment to confirm that MWEs (verbal structures) can be treated in a similar way
  - Results were presented at the 5th Terminology Summit

Enlargement of HUN-LIT proto-dictionary

Augmenting the size of the parallel corpus

Parallel Corpus	Tokens	Translation Units
Lithuanian ORIG	1,765,000	147,158
Lithuanian NEW	**3,544,000**	**262,423**
Hungarian ORIG	2,121,000	147,158
Hungarian NEW	**4,189,000**	**262,423**

Enlargement of HUN-LIT proto-dictionary

Refining the parameters

HUN	LIT	EN	Translation Probability P(tr)	Freq HUN	Freq LIT
gondosan	rūpestingai	'carefully'	**0.22123**	**218**	**118**

- *P(tr)*, *source* and *target lemma frequencies* served as parameters to automatic selection of the best candidates
- Goal:
  - to select the best translation candidates
  - to keep as many translation candidates as possible
- A trade-off between the two objectives has to be found

Enlargement of HUN-LIT proto-dictionary

Refining the parameters

Lemma frequency (LF)	Translational probability P(tr)	Number of translation candidates	Number of evaluated translation candidates
$5 \leq LF < 30$	$P(tr) \geq 0.3$	6713	200
$30 \leq LF < 90$	$P(tr) \geq 0.1$	5181	200
$90 \leq LF < 300$	$P(tr) \geq 0.07$	3401	200
$300 \leq LF$	$P(tr) \geq 0.04$	2725	200

5,521 translation candidates with the original parameters

18020	800

Enlargement of HUN-LIT proto-dictionary

Refining the parameters – Evaluation results

Evaluation range	OK	Useful	OK + Useful	Too special vocabulary	Incorrect	Useless	Number of useful candidates
$5 \leq LF < 30$ $P(tr) \geq 0.3$	40%	24 %	**64%**	7%	29%	**36%**	**4,296**
$30 \leq LF < 90$ $P(tr) \geq 0.1$	59%	21%	**80%**	8,5%	11,5%	**20%**	**4,144**
$90 \leq LF < 300$ $P(tr) \geq 0.07$	75%	14%	**89%**	9%	2%	**11%**	**3,026**
$300 \leq LF$ $P(tr) \geq 0.04$	43.5 %	35%	**78.5%**	9.5%	12%	**21.5%**	**2,139**
							13,605

Creating the French-Dutch Dictionary

Description of the source parallel corpus

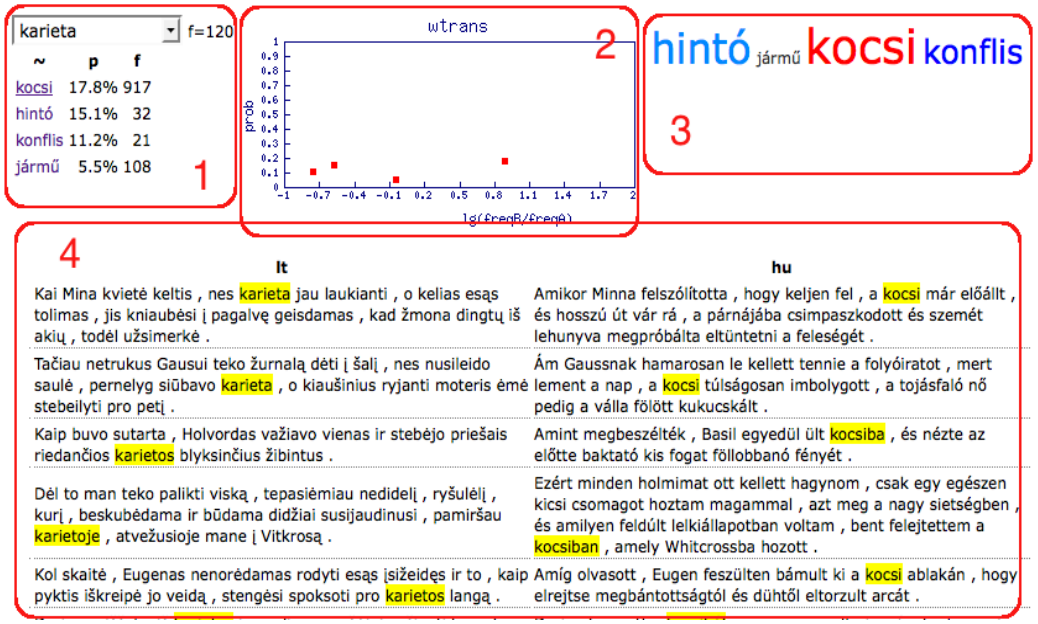
- French-Dutch subcorpus of the Dutch Parallel Corpus (DPC)
  - TLT-Centrale (Macken et al., 2007)
  - FRENCH: 3,606,000 tokens; DUTCH: 3,215,000 tokens
  - BOTH: 186,945 translational units
  - Morphological annotation
  - Various text types
    - literature, journalistic texts, instructive texts, administrative texts, external communication

Creating the French-Dutch Dictionary

Results - with the original parameter setting

- Workflow was the same as in the case of the HUN-LIT language pair
- Results were comparable to that of the HUN-LIT proto-dictionary in terms of precision and recall (7007 translation candidates with the original parameter setting)
- Refinement of the parameters increased considerably the number of the likely translation candidates

Online Versions: Dictionary Browser



- (3) Word cloud:
  - font size mirrors P(tr)
  - colour reflects semantic relation

  between source and target headwords

- (1) Translation candidates are ranked based on their likelihood => most used translation candidates come first
- (2) Plot displays the distribution of translations based on P(tr) and ferquency ratio between the source word and the corresponding translation
- (4) Relevant contexts can be easily listed by clicking on the translation candidate

Link

http://efnilex.nytud.hu/efnilex/

Future plans

- **Multi-word expressions**
  - Our proof-of-concept experiment on verbal structures should be extended to handle **collocations**, too
- Adding **monolingual frequency lists** to compensate for accidental gaps in coverage and to provide a balanced list of lemmas.
- Predicting **semantic relations** (hyponymy, hyperonymy, translational equivalence) between source and target lemmata in the dictionaries

References

- Váradi T., Héja E.: Multilingual Terminology Extraction from Parallel Corpora – A Methodology for the Automatic Extraction of Verbal Structures and their Translation Equivalents. In: *Journal of Hungarian Terminology* (to appear)
- Héja E.: Dictionary Building based on Parallel Corpora and Word Alignment. In: Dykstra, A. and Schoonheim, T., (eds): *Proceedings of the XIV. EURALEX International Congress*, 2010, pp. 341-352.
- Héja E.: The Role of Parallel Corpora in Bilingual Lexicography. In: *Proceedings of the LREC2010 Conference*, La Valletta, Malta, May 2010, pp. 2798-2805.

Thank you for your attention!

http://efnilex.nytud.hu/efnilex/

Outline

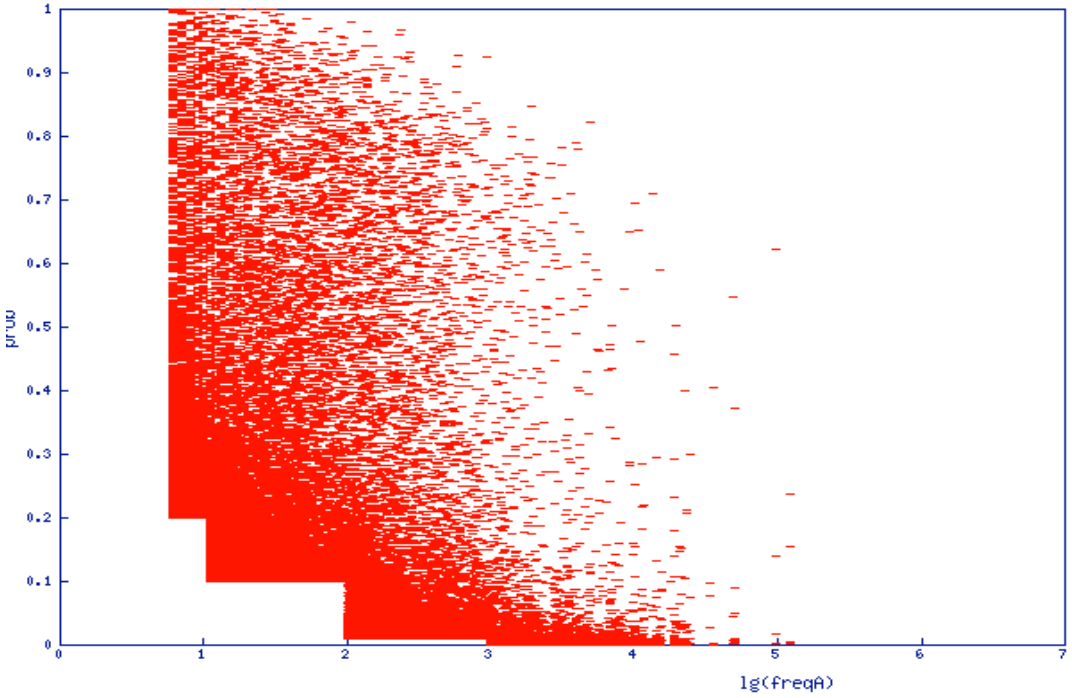
- Background
- Recent activities
  - Increasing the size of the Hungarian-Lithuanian dictionary
  - Augmenting the size of the parallel corpus
  - Fine-tuning the parameters
  - Creating the French-Dutch dictionary
- Online dictionary query system
- Future plans

Advantages II

- *Encoding dictionary*: designed to help people to make utterances in a foreign language ➔ relevant contexts giving hints on the proper use of an expression are particularly important
- Reversing the dictionary is a more straightforward process

Online Versions: Customization

Distribution of HUN-LIT translation candidates



- Customizable in terms of precision and coverage to suit different user needs
  - Novice language learner: reliable translations for basic vocabulary (high precision, low coverage)
  - Professionals: special uses of words, able to select the proper equivalents (low precision, great coverage