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

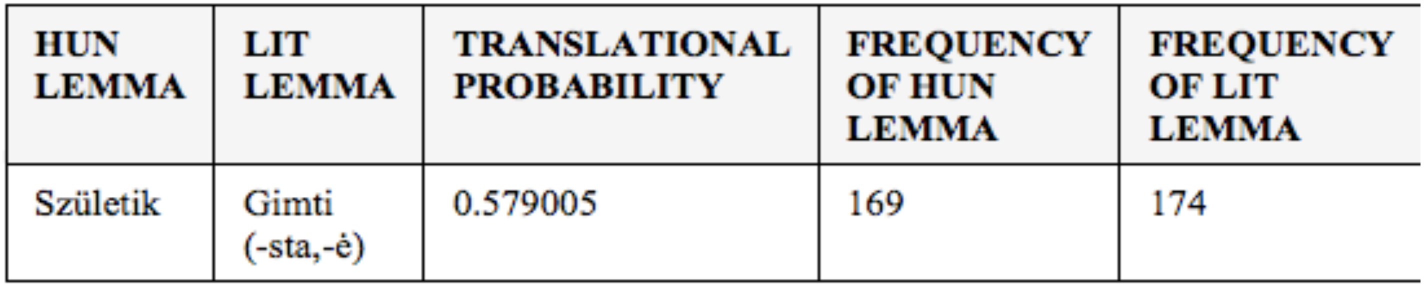
- **Language pairs**:
  - Hungarian – Slovenian
  - Hungarian – Lithuanian

Advantages

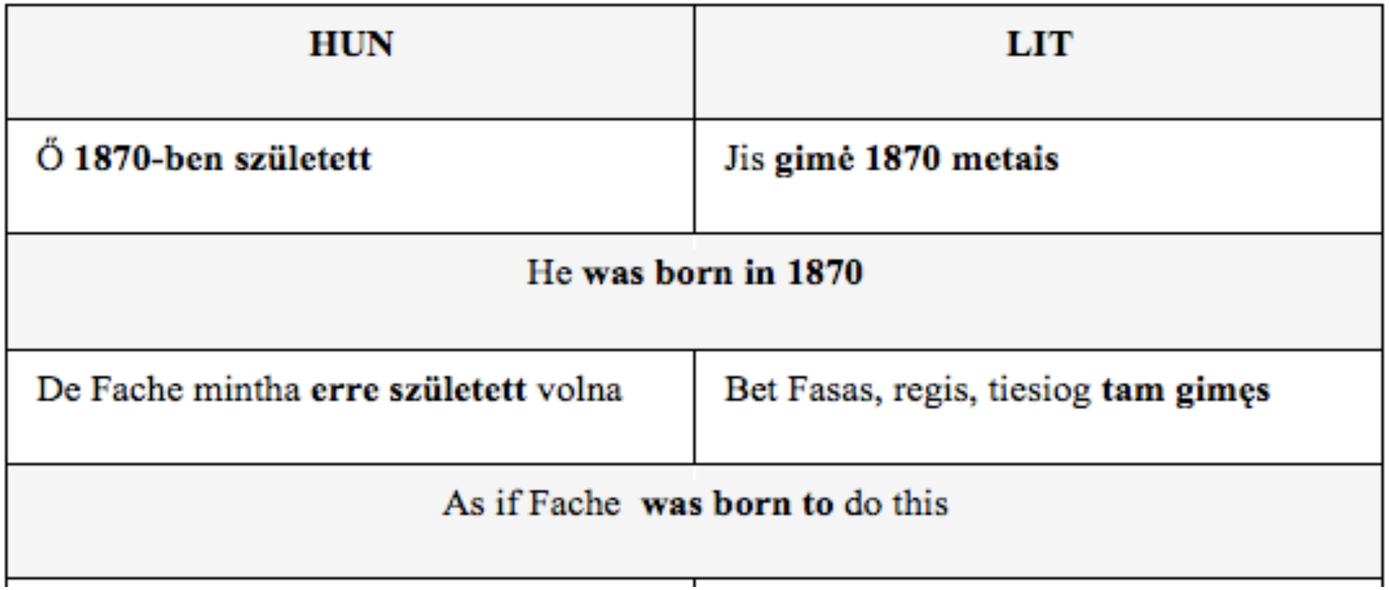
- Parallel corpus => *Corpus-driven technique* to diminish the role of lexicographers' intuition
    - Usage-based, representative translations
    - Clear ranking between more likely and less likely translations
        - Most-used translation equivalents are ranked higher (Example I)
    - Provided 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

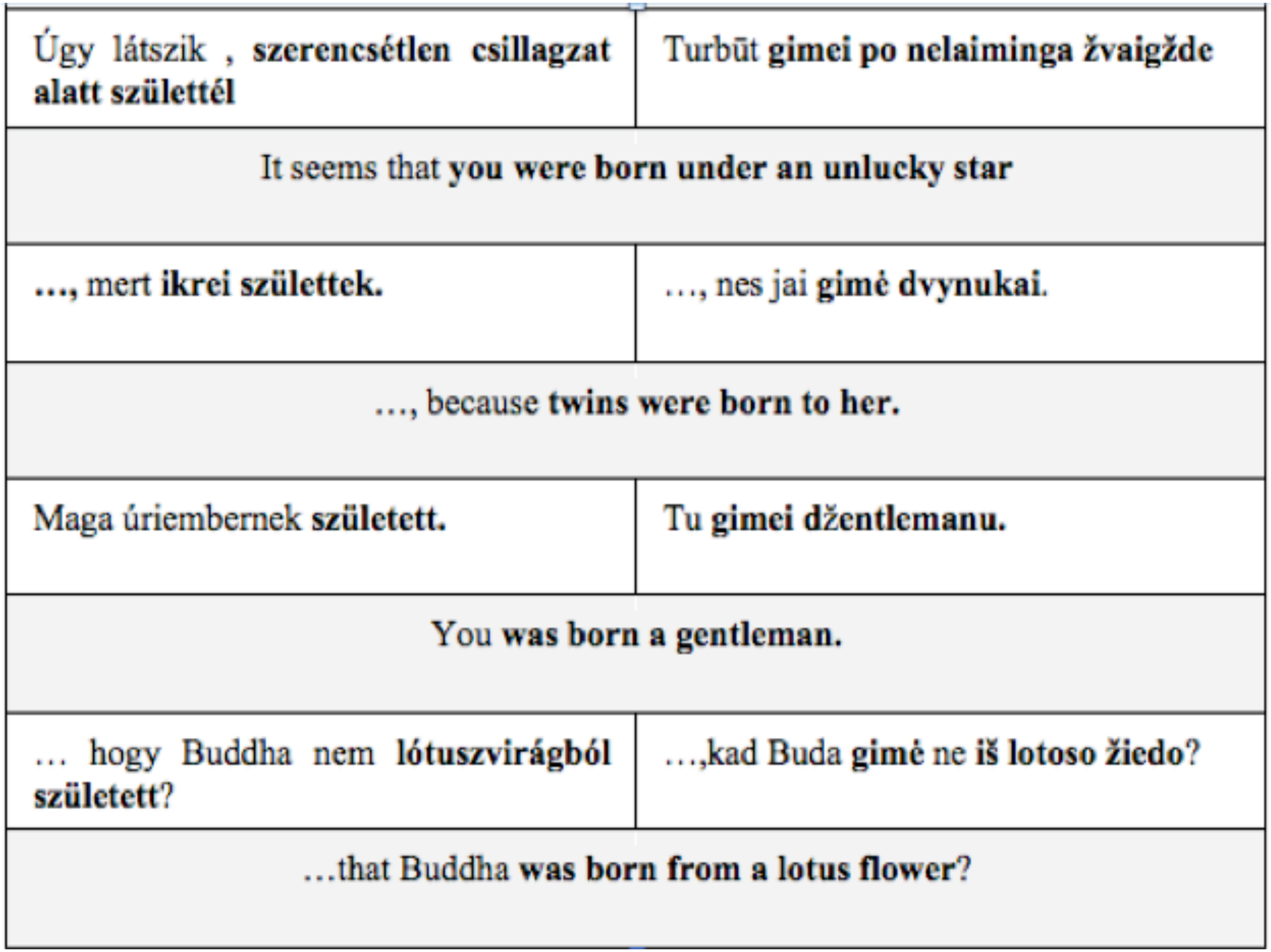


HUN LEMMA	**LIT LEMMA**	**TRANSLATIONAL PROBABILITY**	**FREQUENCY OF HUN LEMMA**	**FREQUENCY OF LIT LEMMA**
Születik	Gimti (-sta,-ė)	0.579005	169	174



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	

Advantages - a Sample



Úgy látszik , **szerencsétlen csillagzat alatt születtél**	Turbūt **gimei po nelaiminga žvaigžde**
It seems that **you were born under an unlucky star**	
…, mert **ikrei születtek.**	…, nes jai **gimė dvynukai**.
…, because **twins were born to her.**	
Maga úriembernek **született.**	Tu **gimei džentlemanu.**
You **was born a gentleman.**	
… hogy Buddha nem **lótuszvirágból született**?	…,kad Buda **gimė** ne **iš lotoso žiedo**?
…that Buddha **was born from a lotus flower**?	

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

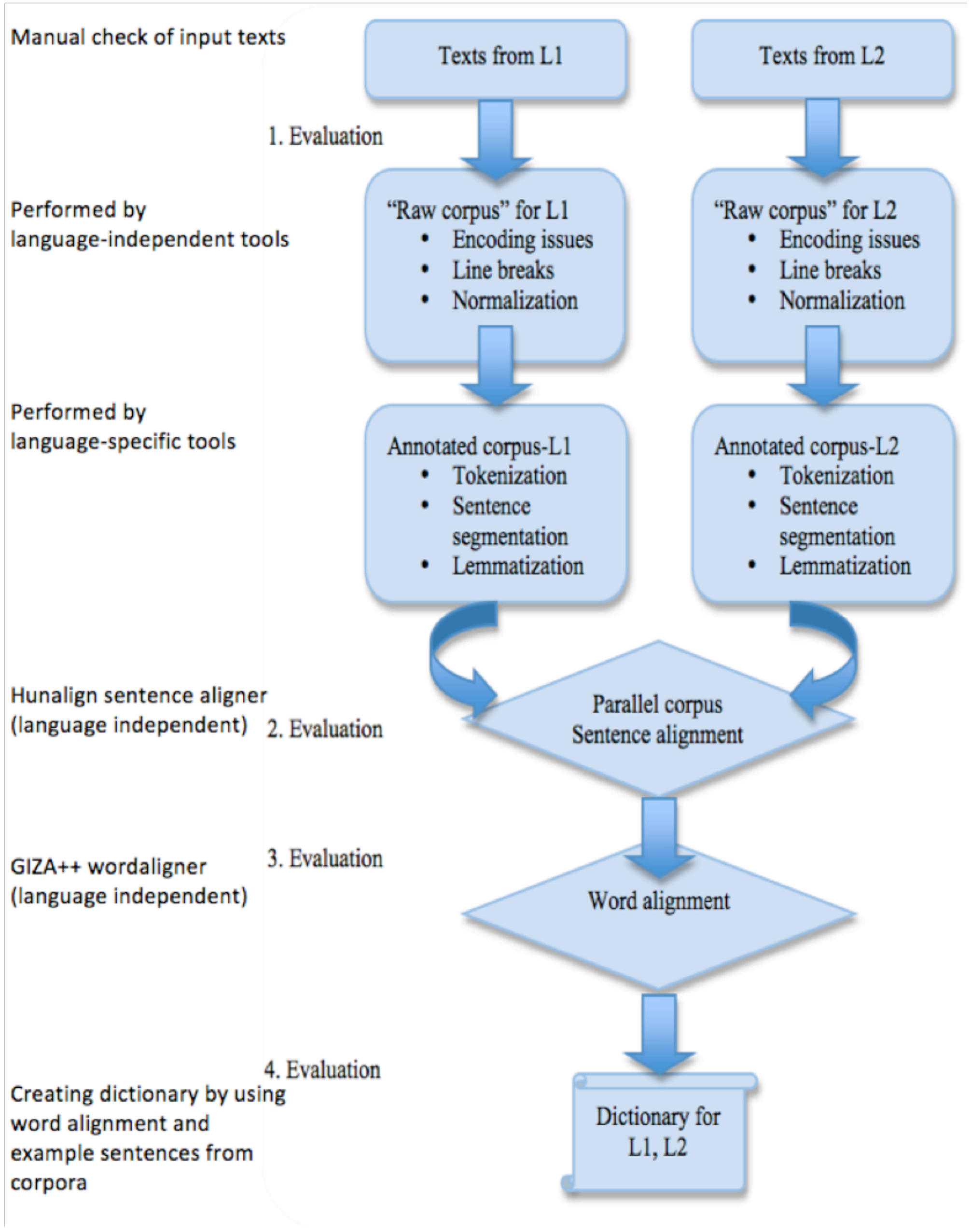
Resources

- 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



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

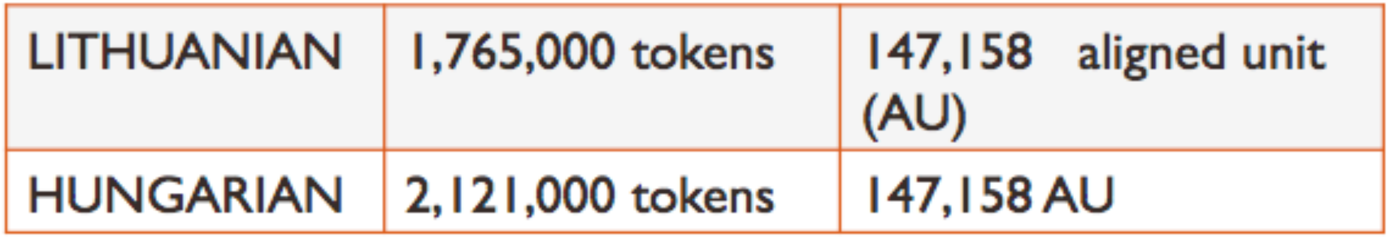
Dictionary for L1, L2

Evaluation Steps

- The quality of the resulting dictionary depends highly on the factors below:
  - Quality of input texts
  - Quality of sentence alignment
  - Quality of word alignment

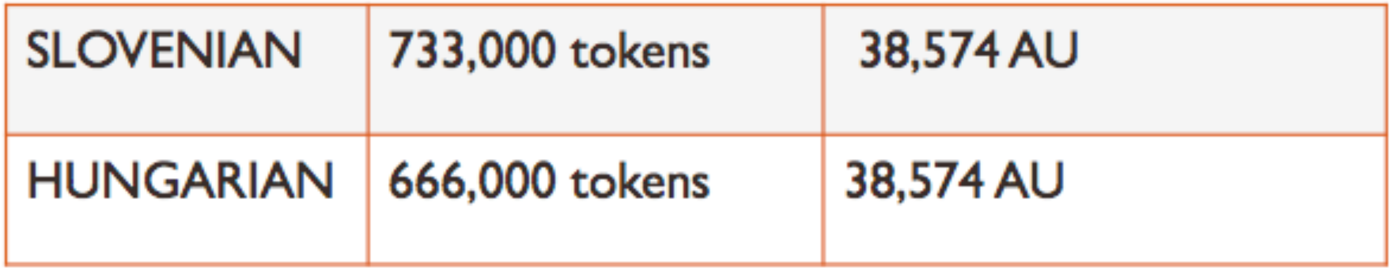
Size of Parallel Corpora

- Lithuanian-Hungarian



LITHUANIAN	1,765,000 tokens	147,158 aligned unit (AU)
HUNGARIAN	2,121,000 tokens	147,158 AU

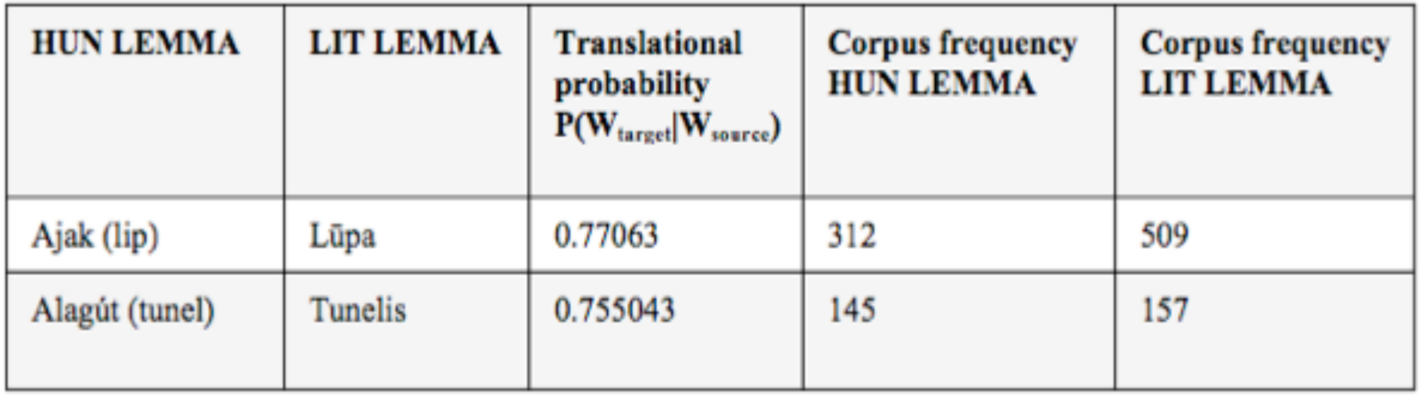
- Slovenian-Hungarian



SLOVENIAN	733,000 tokens	38,574 AU
HUNGARIAN	666,000 tokens	38,574 AU

Most Probable Translation Candidates I

- After word alignment we had the following data at our disposal:



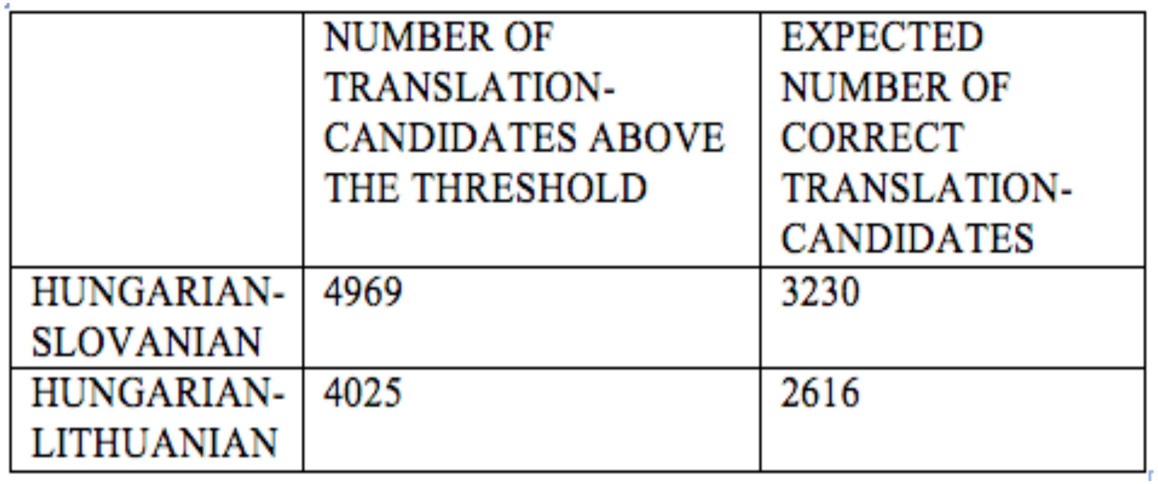
HUN LEMMA	LIT LEMMA	Translational probability $P(W_{target}\|W_{source})$	Corpus frequency HUN LEMMA	Corpus frequency LIT LEMMA
Ajak (lip)	Lūpa	0.77063	312	509
Alagút (tunel)	Tunelis	0.755043	145	157

- *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



	NUMBER OF TRANSLATION-CANDIDATES ABOVE THE THRESHOLD	EXPECTED NUMBER OF CORRECT TRANSLATION-CANDIDATES
HUNGARIAN-SLOVANIAN	4969	3230
HUNGARIAN-LITHUANIAN	4025	2616

Evaluation: Useful Translation Candidates

- Correct translational equivalents
  [gyümölcs – vaisius (fruit)]
- Partially correct translational equivalents => Post editing is needed
  - Improper lemmatization
  - Only partial match in the case of MWEs

    compounds [fo**felügyelő** – vyriausiasis **inspektorius** (chief **inspector**)]

    collocations [**bíborosi** testület – Kardinolų **kolegiją** (cardinal college)]

- Looser semantic relation (e.g. hypernymy)
  [**lúdtoll** (literally**:** goose-feather) – **plunksna** (literally: feather, pen)]
  intended meaning in both cases: *quill pen*

Evaluation: Useless Translation Candidates

- Irrelevant vocabulary (e.g. recurrent proper names) [**Abdul** – **Abdulas**]

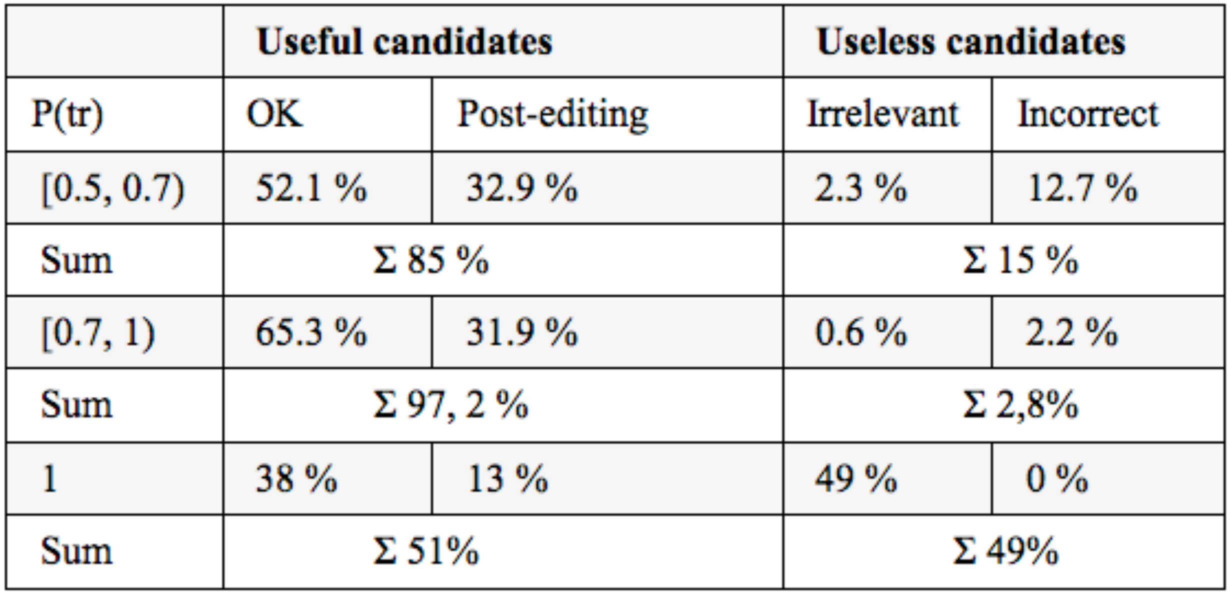
- Incorrect translation candidates
  - Usually due to the loose translations of texts

Evaluation – Data

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

- Evaluation intervals:
  - $0.5 \leq P(w_{target} | w_{source}) < 0.7$
  - $0.7 \leq P(w_{target} | w_{source}) < 1$
  - $P(w_{target} | w_{source}) = 1$

Results



	Useful candidates		Useless candidates	
P(tr)	OK	Post-editing	Irrelevant	Incorrect
[0.5, 0.7)	52.1 %	32.9 %	2.3 %	12.7 %
Sum	Σ 85 %		Σ 15 %	
[0.7, 1)	65.3 %	31.9 %	0.6 %	2.2 %
Sum	Σ 97, 2 %		Σ 2,8%	
1	38 %	13 %	49 %	0 %
Sum	Σ 51%		Σ 49%	

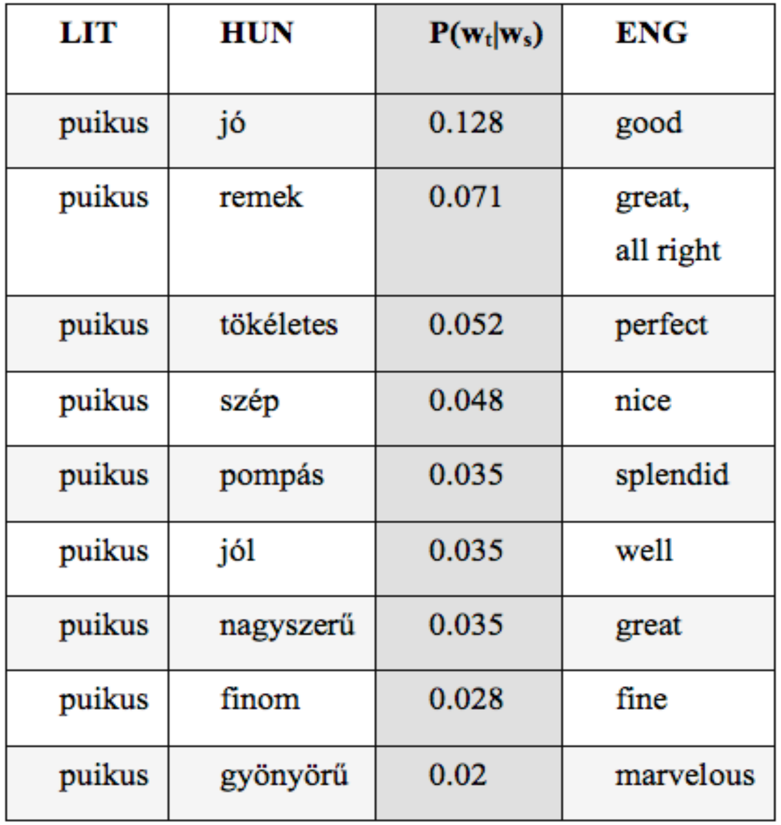
- Proportion of incorrect translation pairs is low
- **85 %** of translation pairs are *useful* in the 1. probability range
- **97,2 %** of translation pairs are *useful* in the 2. range
- P(tr)=1 produces the lowest proportion of useful candidates and the highest ratio of irrelevant pairs

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 → 0.02)

Related Meanings – Example I



LIT	HUN	$P(w_t\|w_s)$	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.

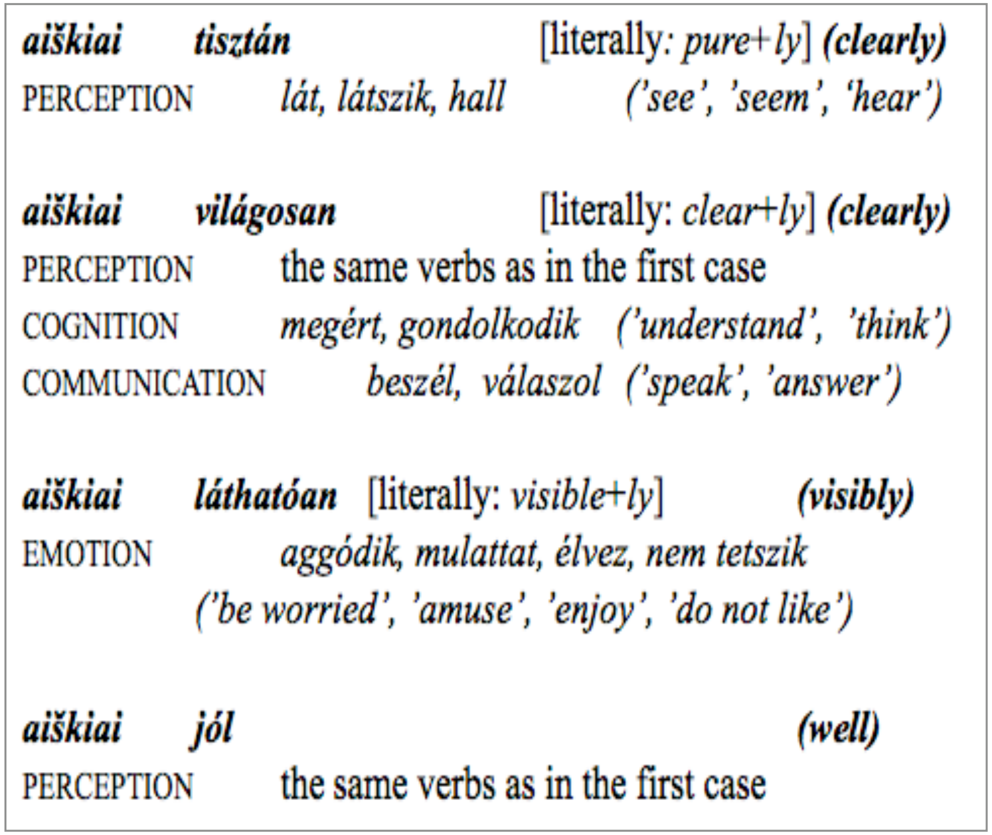
-**Remek** – válaszolta a hang. (-**All right** – the voice answered)

-**Puikus** darbas.

-**Szép** munka volt. (-**Good** job)

Related Meanings – Example II

- Use in the creation of encoding dictionaries



aiškiai ***tisztán*** [literally: *pure+ly*] ***(clearly)***

PERCEPTION *lát, látszik, hall* *('see', 'seem', 'hear')*

aiškiai ***világosan*** [literally: *clear+ly*] ***(clearly)***

PERCEPTION the same verbs as in the first case

COGNITION *megért, gondolkodik* *('understand', 'think')*

COMMUNICATION *beszél, válaszol* *('speak', 'answer')*

aiškiai ***láthatóan*** [literally: *visible+ly*] ***(visibly)***

EMOTION *aggódik, mulattat, élvez, nem tetszik*
('be worried', 'amuse', 'enjoy', 'do not like')

aiškiai ***jól*** ***(well)***

PERCEPTION the same verbs as in the first case

- *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!