Workflow

1. Overview:

(i) **Languages**: At the meeting in the Hague (14/01/2009) we have decided that at the annual EFNIL meeting (04/11/2009) automatically generated dictionaries should be presented for two language pairs:

- Hungarian - Lithuanian (languages with fewer resources)
- Dutch - French (languages with richer resources)

(ii) **Expectations**: At the beginning of the project, EFNILEX set the following expectations toward the dictionaries:

- They should cover everyday language usage
- A medium sized dictionary of 20.000-45.000 entries was suggested (the size depending also on the available resources and tools for the languages)

(iii) **Methodology:**

- Statistical word alignment
- On parallel corpora

(iv) **Corresponding main tasks:**

- Creation of parallel corpora (if it does not exist for the given language pair)
- Performing word alignment
- Creation the dictionaries based on the word alignment
- Evaluation

(v) **Structure of the presentation**

Based on the main tasks above this presentation is organized as follows:

1. Collecting resources and tools
2. The building process
3. Evaluation
4. Presentation of the prototypes
5. TODOs

2. Collecting resources and tools

(i) What to collect?

- **Resources**: Based on the preliminary expectations toward the dictionaries, we have estimated that a 10.000.000-token corpus for each language would be sufficient for our purposes.
- **Tools**: Only language-dependent tools need to be collected for each language. I. e.:
  - Sentence splitters
  - Tokenizers
  - Lemmatizers
  - Taggers

(ii) Collecting parallel texts

- Dutch-French (Annemieke's recent letter)
- Lithuanian-Hungarian, Slovenian-Hungarian
- For medium-density languages collecting direct translations does not seem a viable approach.

  - Lithuanian: 0 tokens
  - Slovenian (through contacting several translators, publishers of books and journals, the Slovenian Television, etc., downloading a few bilingual web pages): ~750.000 tokens.

- Instead of direct translations we have decided to collect translations from a third language (e. g. English, French, German, etc.)
  - Parallel web pages from the web. EU news[1] for all the languages in question (~200.000 tokens per language).
  - Collecting mainly pieces of literature from the web (especially for Hungarian, where the national digital archives (MEK, DIA) proved to be rich resources of fictions).
  - Contact the organizations which have created the national corpora for Lithuanian and Slovenian so that we could make use of their texts.

1. *Vytautas Magnus University, Centre of Computational Linguistics*,

[1] http://ec.europa.eu/news/archives_en.htm

- *Rūta Marcinkevičienė* (texts from Lithuanian National Corpora : fiction: ~150, non-fiction: ~40)
- *Andrius Utka* (texts from the Lithuanian-English parallel corpora: fiction: ~65, non-fiction: ~48)

2. *Slovanian FIDA corpus*: *Tomaz Erjavez* has sent a list of the texts from FIDA corpus (they gave us texts in return for their Hungarian counterparts).

- If we have a reasonable amount of general-domain text (~9.000.000 tokens) we can add more domain-specific but easily available texts (legal or religious).

(iii) Collecting language dependent tools

Our contacts provided us with the necessary tool-chains. These tool-chains comprise all the language dependent tools we need.

- LIT: the Lithuanian Centre of Computational Linguistics performs the annotation of required texts.
- SLO: tool-chain is available on the web: http://nl.ijs.si/jos/analyse/ The tool-chain performs all the tasks needed, unfortunately only one file at a time is processed.
- HUN: HNC tool-chain to analyze great amount of texts fully automatically.

(iv) Results

Slo-Hun	910.300 token	55.300 sent	794.800 token	47.300 sent
Lit-Hun (annotated)	1812338 token	179467 sent	2.301.360 token	166.765 sent

LIT:

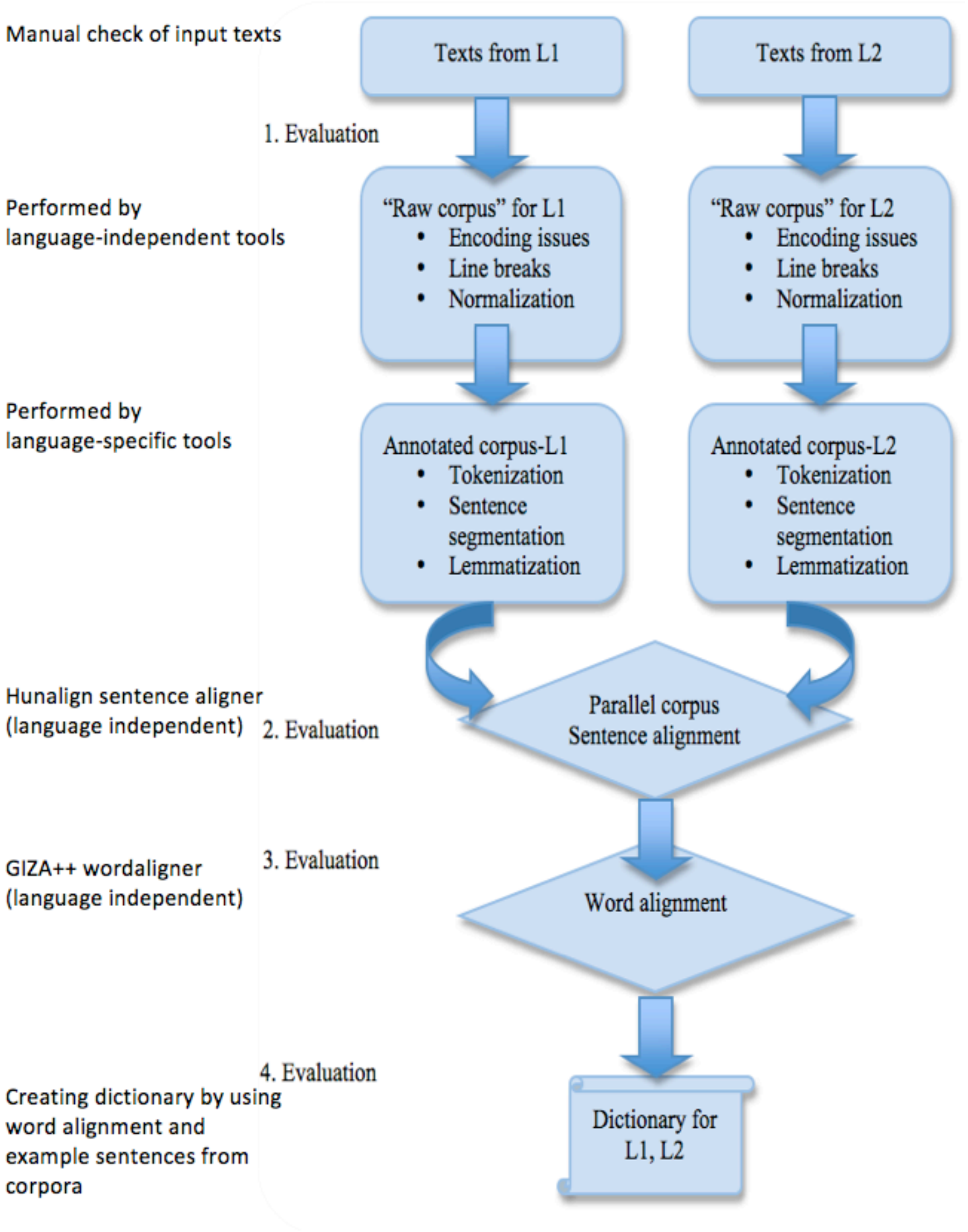
- One-million token is under processing right now.
- TODO: Trying to find more Hungarian equivalents through contacting publishers.

SLO:

- TODO: The same might be done also for Slovenian.

3. The dictionary building process

The building process consists of the following steps:



(i) In several cases, there is an evaluative step between two stages of the building process, since the quality of the dictionary seems to be highly dependent on:

- Noisiness of the input texts
- The quality of the output of the language dependent tools.
- The output of the language dependent tools (mainly lemmatization) needs further evaluation.

(ii) Sentence alignment and word alignment assign confidence values to the generated sentence and word pairs, respectively. We can filter out the most probable pairs based on these values.

(iii) We have built dictionaries based on 3 different input corpora:

(a) Slovenian and Hungarian EU-news corpora (~200.000 token/each).

- The manual evaluation of the produced parallel corpora showed the importance of using cleaned input texts to achieve good sentence alignment.
  - Filter out untranslated English texts
  - Pay attention to html entities, links, etc.
- Based on manual evaluation of translation pairs (with translation probabilities between 0,7 and 1) we concluded that there is a considerable decrease in the correctness of translation if both type occurs less than five times in the corpora.

Frequency of both tokens	Right translations 0,7 <= P(tr) <= 1	Right translations 0,5 <= P(tr) < 0,7	Right translations P(tr) < 0,5
>= 3	60%		The proportion of the right translation pairs is dropping considerably independently of the occurence frequency
>= 4	70%		
>= 5	76%	57%	
> 10	76%		
< 5	25%	14%	

(b) Slovenian corpus consisting of all the texts available:

- Manual preprocessing of input texts (missing or untranslated parts)
- Manual evaluation of sentence alignment yielded the result that using a confidence value equal or higher than 0,1 gives a reliable sentence alignment.

- Filtering out these sentences has been modified the size of the original corpora:

Size of the original corpora (SLO)	910.300 tokens	55.300 sent
Aligned sentences with 0.1 conf value or more (SLO)	734.700 tokens	**38.574 sent**
Size of the original corpora (HUN)	794.800 tokens	47.300 sent
Aligned sentences with 0.1 conf value or more (HUN)	667.000 tokens	**38.574 sent**

- While creating the dictionary we have considered also the thresholds from the first experiment:
  - We prepared the dictionary with translation pairs where each of the tokens has a frequency greater than 5.
  - To keep as much translation pairs as possible we set the translation probability to 0.2.

- Size of the dictionary:

HUN-SLO word pairs: P(tr) > 0,2 and Freq(HUN) > 4 and Freq(SLO) > 4	11.700 word pairs

(c) Lithuanian corpus consisting of all annotated texts:

- Manual check of the input texts
- For the sentence alignment we have applied the threshold gained from the second Slovenian experiment:
  - We have filtered out aligned sentences with a confidence value less than > 0,1
- The size of the original corpora has been modified according to the values in the table below:

Size of the original corpora (LIT)	1.812.338 tokens	179.467 sent
Aligned sentences with 0.1 conf value or more (LIT)	1.501. 400 tokens	**122.900 sent**
Size of the original corpora (HUN)	2.301.360 tokens	166.765 sent
Aligned sentences with 0.1 conf value or more (HUN)	1.801.300 tokens	**122.900 sent**

- Also the thresholds from the first experiment have been made use of while creating the dictionary:

- Each token in the dictionary has to occur at least 5 times in the corpora. Translation pairs where each of the tokens has a frequency greater than 5.
- To keep as many translation pairs as possible instead of 0.5 we set the translation probability to 0.2.

– Size of the dictionary:

HUN-LIT word pairs: P(tr) > 0,2 and Freq(HUN) > 4 and Freq(LIT) > 4	20.300 word pairs

4. Evaluation of the dictionaries

(i) **Goal**: Since producing perfect dictionaries fully automatically with the state of the art methods and tools seems to be completely impossible, our basic objective is to support human lexicographers by computational linguistic means as much as possible, so that the invested effort could be decreased.

- Instead of producing dictionary with only proper translations, our intention is to gather as many translations as possible (this entails with the inclusion of wrong translations)

(ii) **Precision:** Correct translation pairs / all the translation pairs we have generated

- As a first approach we estimated that a 0.8 precision value might be sufficient.
- Based on the greater amount of input data we made the presupposition that with an appropriate frequency of the types (5), even a law transition probability (0.2) might yield appropriate results.
- Although evaluation of the dictionaries is an extremely time consuming task and also needs considerable expertise, it cannot be omitted, especially in this early stage of the project.
  - SLO: a PhD-student of Tamás.
  - LIT: ???

(iii) **Recall:** right translation pairs/all the existing translation pairs between the two languages

- Although recall (or the coverage) of a dictionary is quite an important feature of it, until now we have not found a paper which describes an exact method to measure it.

- Comparing the coverage with that of existing dictionaries?
- With the most frequent words in a reference corpus?

(iv) **Manual Evaluation of the HUN-SLO dictionary:**

Precision:

- It was performed on the first 200 entries (*freq(HUN) > 4, freq(SLO) > 4, P(tr) > 0.2*) with the following results:
  - Correct translations: *57,7%*
  - Multiword expressions: *10,5%*
  - Since the applied word alignment method enables only 1-to-1 correspondences, the handling of MWEs poses a general problem for us.
  - Typical cases of 1-to-more mappings in Slovenian:
    - Compound words in HUN
    - Reflexive verbs in SLO

Recall:

- By comparing it with Elizabeta Bernjak's Hungarian-Slovenian dictionary (Cankarjeva, 1995)
- The 200th word in the automatically generated dictionary is 'Ana'. In Bernjak's dictionary there are 530 entries before this string. Considering the fact that a real dictionary comprises headwords instead of word pairs, the difference might be even greater.
- On the other hand, the size and representativeness of the parallel corpora has direct influence on the amount and quality of the translation pairs. Thus, adding more texts could easily improve the recall.

Examples:

- The data driven generation method re-interprets the notion of translational equivalence. By doing so, it supplies correct translation possibilities, which might remain hidden to the lexicographers intuitions.
- Although automatically generated dictionaries do not give explicit grammatical information on the entries, the parallel example sentences might yield useful clues on the usage of the words.

(1) Automatically generated dictionary:
Agyonver (*beat to death*) – *razmazati (szétken - smash)*

Although, there is no dictionary containing these words as direct translations, based on the contexts they can be translational equivalents.

(2) In usual dictionaries, translations have the same part-of-speeches. In these cases, we probably do not have to pay attention to this expectation, since from the context the usage of the word is again obvious.

In Bernjak's dictionary:
afrikai I. –ul *mn* áfriški; **II.** –t, -ak, -ja *fn* Afričán, (nő) Afričánka
African male female

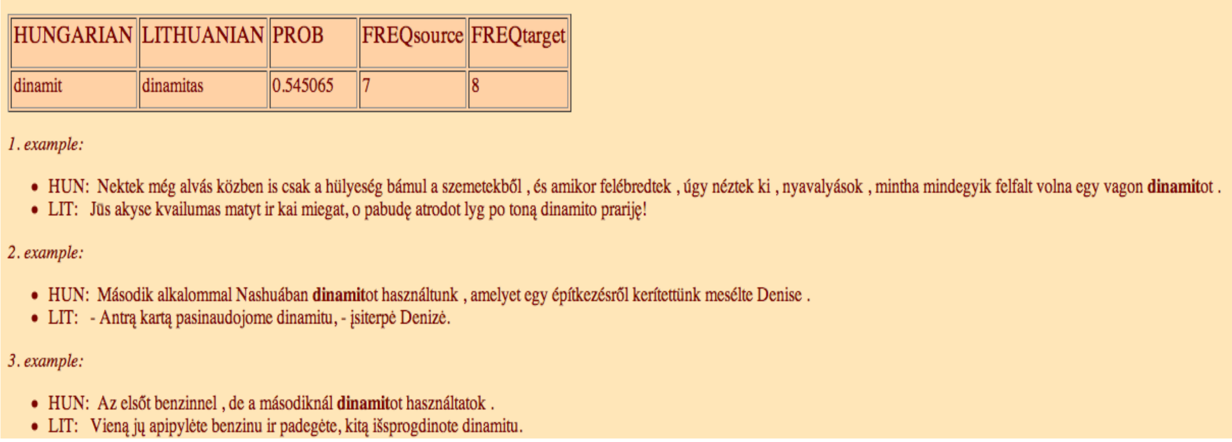
Our dictionary lists the noun *Africa* besides the adjective. And indeed, the examples illustrate that the Hungarian adjective *Afrika* is translated several times into Slovenian as a noun.

HUN:
A győztesek jutalma afrikai utazás,
The winners' prize is an African journey.

SLO:
Nagrajenci bodo obiskali Afriko.
The winners visit Africa.

5. Prototypes

Hungarian-Slovanian:	http://corpus.nytud.hu/people/eheja/efnilex/slo_dic_final.zip
Hungarian-Lithuanian:	http://corpus.nytud.hu/people/eheja/efnilex/lit_dic_final.zip



HUNGARIAN	LITHUANIAN	PROB	FREQsource	FREQtarget
dinamit	dinamitas	0.545065	7	8

1. example:

- HUN: Nektek még alvás közben is csak a hülyeség bámul a szemetekből , és amikor felébredtek , úgy néztek ki , nyavalyások , mintha mindegyik felfalt volna egy vagon **dinamit**ot .
- LIT: Jūs akyse kvailumas matyt ir kai miegat, o pabudę atrodot lyg po toną dinamito prariję!

2. example:

- HUN: Második alkalommal Nashuában **dinamit**ot használtunk , amelyet egy építkezésről kerítettünk mesélte Denise .
- LIT: - Antrą kartą pasinaudojome dinamitu, - įsiterpė Denizė.

3. example:

- HUN: Az elsőt benzinnel , de a másodiknál **dinamit**ot használtatok .
- LIT: Vieną jų apipylėte benzinu ir padegėte, kitą išsprogdinote dinamitu.

6. TODOs

i. Increase the size of the corpora
ii. Build a more robust architecture
iii. Extension of word mappings to MWEs
iv. Evaluation