Efficient Exploration of Translation Variants in Large Multiparallel Corpora Using a Relational Database

Johannes Graën Simon Clematide Martin Volk

Institute of Computational Linguistics
University of Zurich, Switzerland

28th May, 2016

Outline

Outline

Motivation

- We identified several types of online parallel corpus query systems[1],
- some of which address the interested public (i.e. non-linguists): Glosbe[2], Linguee[3], Tradooit[4]...
- These systems provide ad-hoc searches with free input instead of a formal corpus query language.

[1] Volk, Graën, and Callegaro 2014.
[2] `https://glosbe.com/`
[3] `http://www.linguee.com/`
[4] `http://www.tradooit.com/`

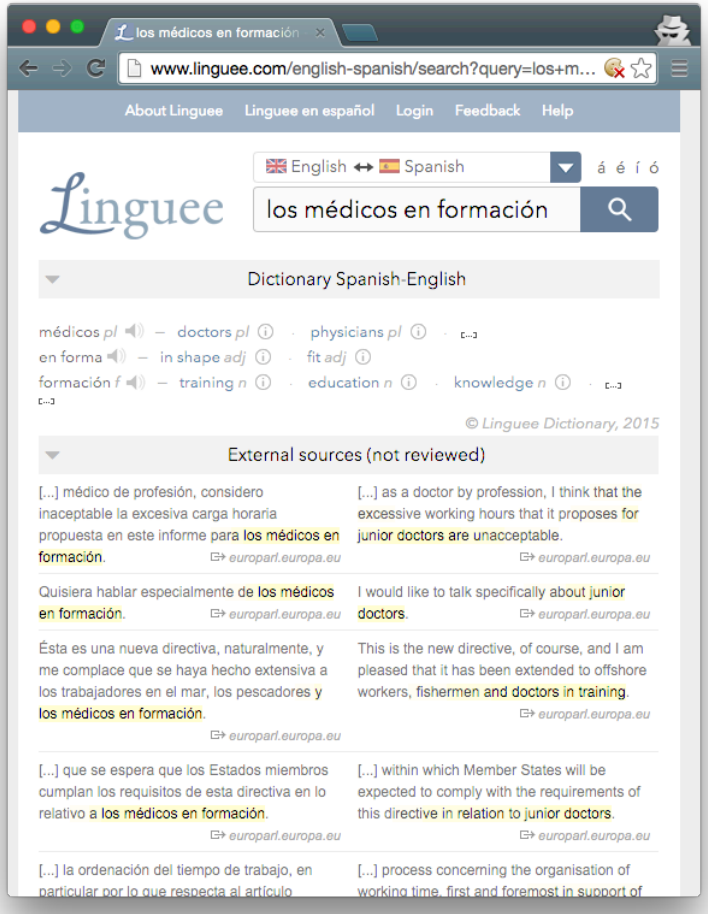
Linguee

For a pair of languages, Linguee

- provides dictionary information,
- lists parallel sentences where the search words appear in one of the languages,
- highlights the search words in the source language and the corresponding words in the target language.

Linguee


Linguee | Dictionary for Ge
www.linguee.com/?chooseDomain=1
About Linguee
Linguee auf Deutsch
Login
Feedback
Help
facebook
twitter
google +1
Linguee
English-German Dictionary.
Search 1,000,000,000 translations.
English ↔ Spanish
English ↔ German
English → German
German → English
English ↔ Portuguese
English ↔ Spanish
English ↔ French
English ↔ Italian
English ↔ Russian
English ↔ Japanese
ä ö ü ß
languages


los médicos en formación
www.linguee.com/english-spanish/search?query=los+m...
About Linguee
Linguee en español
Login
Feedback
Help
English ↔ Spanish
á é í ó
Linguee
los médicos en formación
Dictionary Spanish-English
médicos pl — doctors pl · physicians pl
en forma — in shape adj · fit adj
formación f — training n · education n · knowledge n
© Linguee Dictionary, 2015
External sources (not reviewed)
[...] médico de profesión, considero inaceptable la excesiva carga horaria propuesta en este informe para los médicos en formación.
europarl.europa.eu
[...] as a doctor by profession, I think that the excessive working hours that it proposes for junior doctors are unacceptable.
europarl.europa.eu
Quisiera hablar especialmente de los médicos en formación.
europarl.europa.eu
I would like to talk specifically about junior doctors.
europarl.europa.eu
Ésta es una nueva directiva, naturalmente, y me complace que se haya hecho extensiva a los trabajadores en el mar, los pescadores y los médicos en formación.
europarl.europa.eu
This is the new directive, of course, and I am pleased that it has been extended to offshore workers, fishermen and doctors in training.
europarl.europa.eu
[...] que se espera que los Estados miembros cumplan los requisitos de esta directiva en lo relativo a los médicos en formación.
europarl.europa.eu
[...] within which Member States will be expected to comply with the requirements of this directive in relation to junior doctors.
europarl.europa.eu
[...] la ordenación del tiempo de trabajo, en particular por lo que respecta al artículo
[...] process concerning the organisation of working time, first and foremost in support of

Outline

Concept

Multilingwis

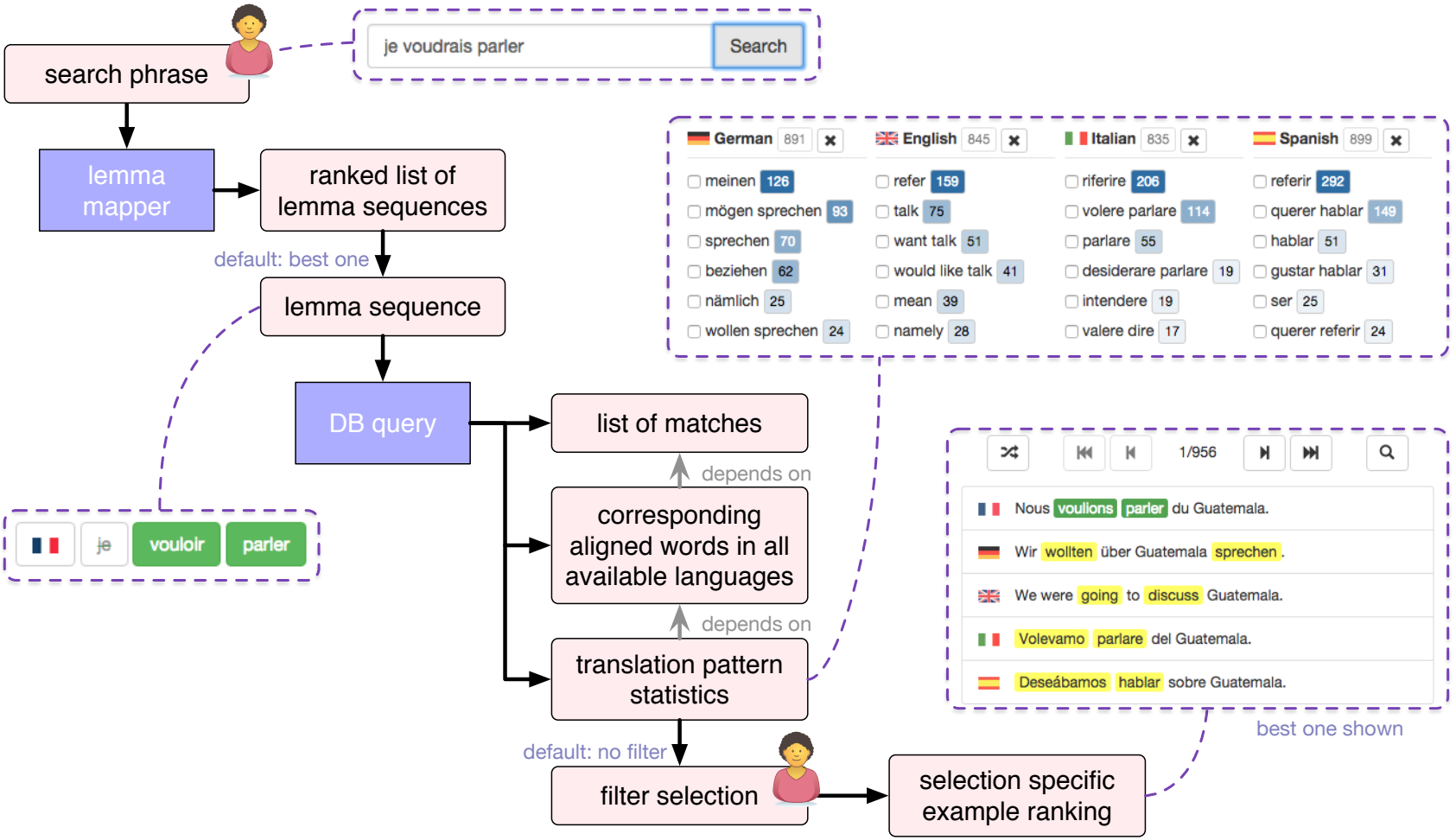
Multilingual Word Information System

We designed Multilingwis to be a corpus exploration system that

- allows for similar ad-hoc searches
- shows the distribution of translation variants
- offers a reverse search for each of those variants
- provides examples with translation equivalents marked

Another important aspect: Prompt responses!

Worflow


search phrase
je voudrais parler
Search
lemma mapper
ranked list of lemma sequences
default: best one
lemma sequence
DB query
list of matches
depends on
corresponding aligned words in all available languages
depends on
translation pattern statistics
default: no filter
filter selection
selection specific example ranking
German 891
English 845
Italian 835
Spanish 899
meinen 126
mögen sprechen 93
sprechen 70
beziehen 62
nämlich 25
wollen sprechen 24
refer 159
talk 75
want talk 51
would like talk 41
mean 39
namely 28
riferire 206
volere parlare 114
parlare 55
desiderare parlare 19
intendere 19
valere dire 17
referir 292
querer hablar 149
hablar 51
gustar hablar 31
ser 25
querer referir 24
je
vouloir
parler
1/956
Nous voulions parler du Guatemala.
Wir wollten über Guatemala sprechen.
We were going to discuss Guatemala.
Volevamo parlare del Guatemala.
Deseábamos hablar sobre Guatemala.
best one shown

Worflow

- A user types in a word or expression.
- The input gets lemmatized and function words are removed.
- The database performs a search for the given sequence of lemmas, where up to 3 function words are allowed in between each two content words.
- It then looks up the alignments, i.e. translation equivalents, for all hits and aggregates them to a frequency distribution of translation variants.
- The overall best example is determined based on shortness and displayed together with the translation variant distribution.

Demo

`http://pub.cl.uzh.ch/purl/multilingwis`

Outline

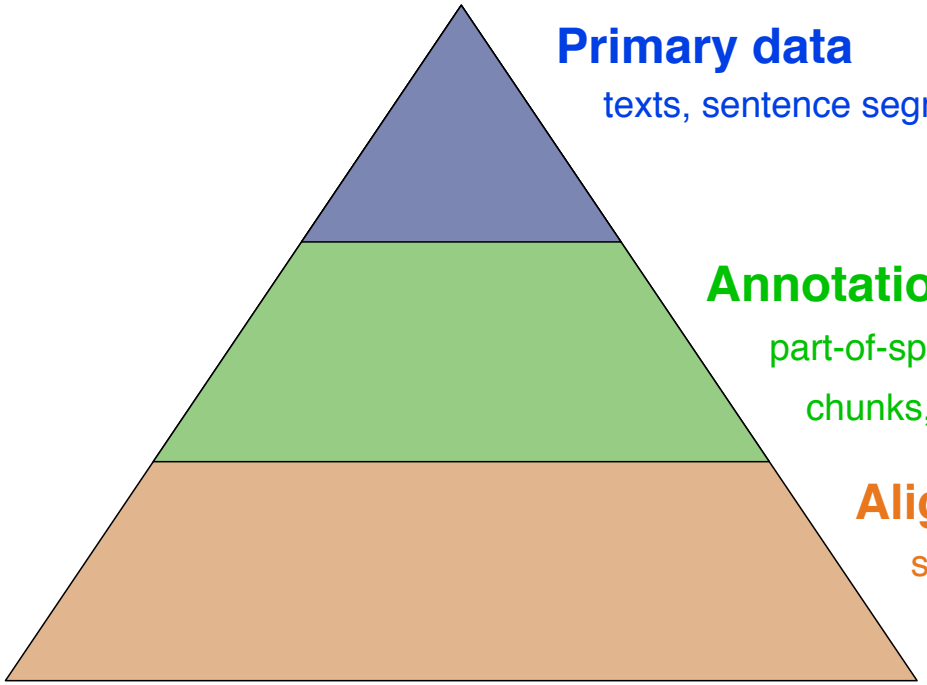
Our Corpus

Version 3

- 146.652 speaker turns from **Europarl/CoStEP**[5] in five languages: English, French, German, Italian and Spanish
- Pipeline:
  - tokenization, part-of-speech tagging and lemmatization with the **TreeTagger** and its featured language models
  - tag mapping to universal part-of-speechtags (uPoS)
  - rule-based sentence segmentation
  - pairwise sentence alignment with **hunalign**
  - pairwise word alignment with **Giza++** based on lemmas ($\equiv$ word form, if no lemma assigned) of content words (ADJ, ADV, NOUN or VERB)

[5] Graën, Batinic, and Volk 2014.

Layout



Database-driven Corpus[6]

all these layers are represented as attributes and relations in a relational database management system (PostgreSQL)

[6] Graën and Clematide 2015.

Figures

- 22 m content words per language
- 1.7 m sentences per language
- 16 m pairwise sentence alignments
- 434 m pairwise content word alignments

Language	Tokens	Types	w/ Lemma	Lemma Ratio
English	43 m	127.105	73.250	57.6 %
French	47 m	142.898	83.937	58.7 %
German	41 m	367.159	174.885	47.6 %
Italian	43 m	181.478	108.147	59.6 %
Spanish	45 m	175.817	75.187	42.8 %

Outline

Efficient Retrieval

Materialized Views and Composite Indexes

For efficient retrieval, we built

- a materialized view on lemmas and relevant foreign keys,
  - only relevant attributes
  - all in a single view
- composite index over all columns starting with the lemma (7.3 GB for 220 million rows),
- another composite index on symmetrized view of word alignments (9.0 GB for 418 million single word alignments).
  - null alignments skipped
  - 'union' symmetrization method for better recall

Efficient Retrieval

Regular B-Tree Index

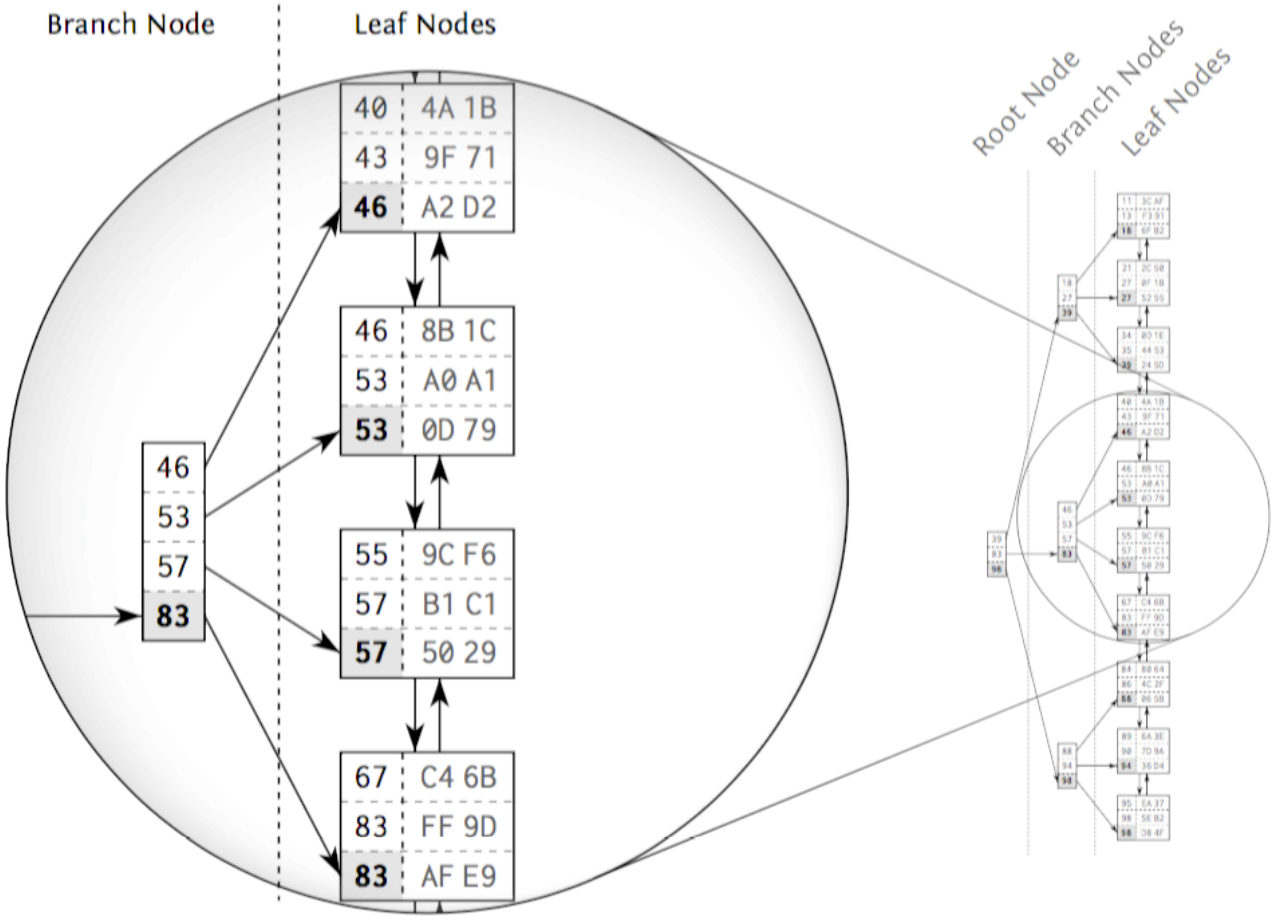


Diagram from `http://use-the-index-luke.com/`

Query

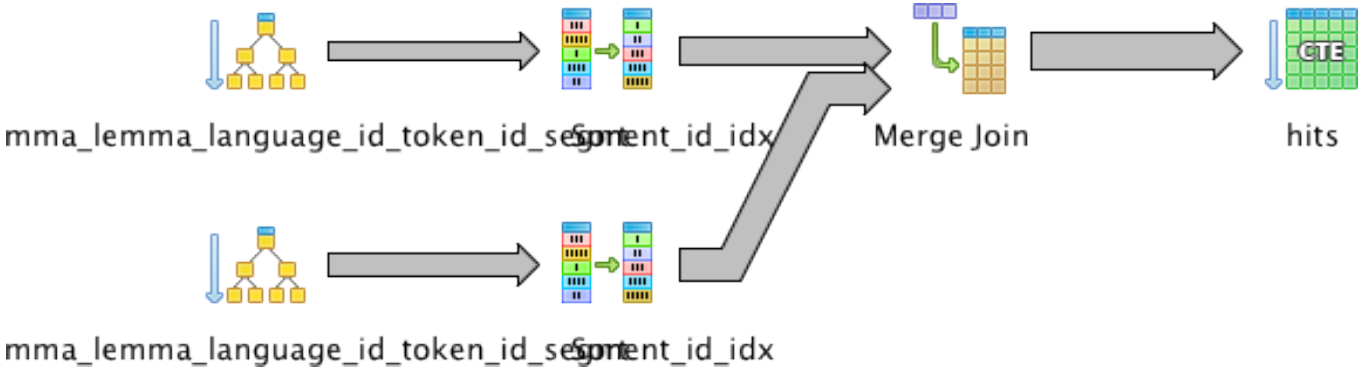
The query function

- scans the lemma index to identify all hits,
- retrieves translation equivalents by intersecting hits with the alignment index,
- joins the lemmas of all aligned tokens and aggregates frequencies of lemma sequences (translation variants).

A particular search function is responsible for each count of source lemmas, allowing for pre-planned queries.

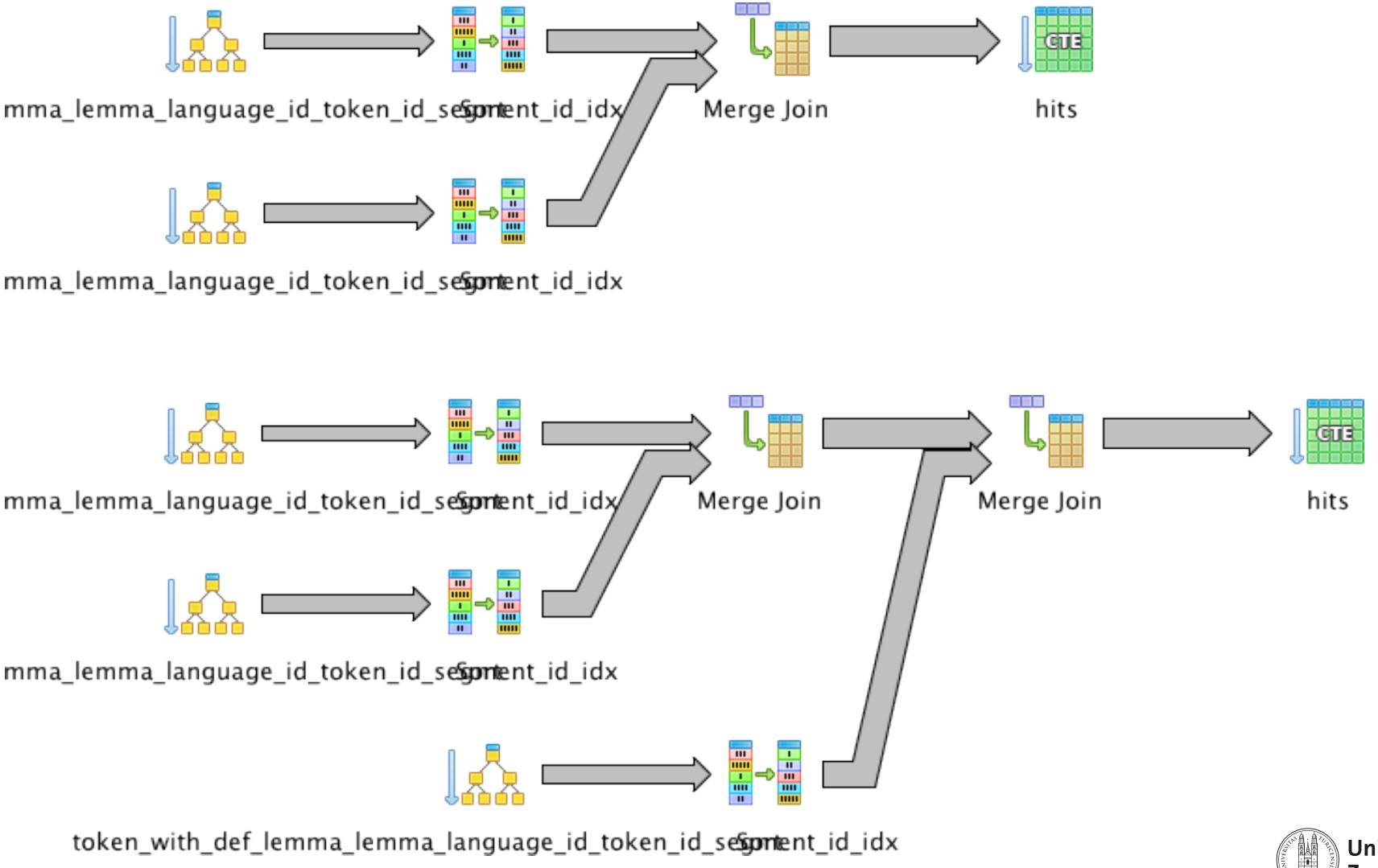
Identifying Hits – Query Plan

for Lists of 2 and 3 Lemmas



Identifying Hits – Query Plan

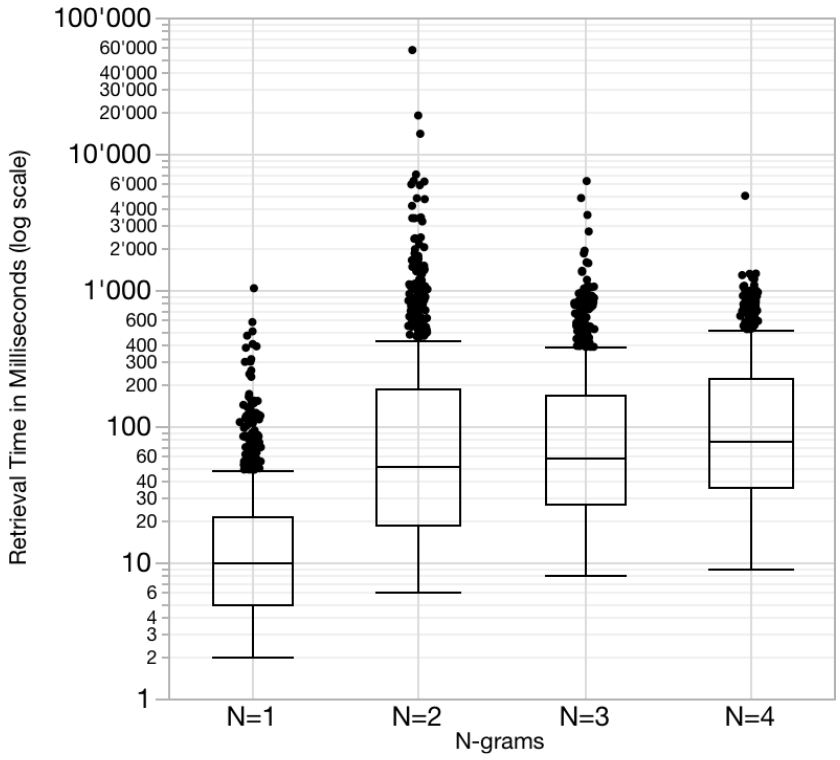
for Lists of 2 and 3 Lemmas



Outline

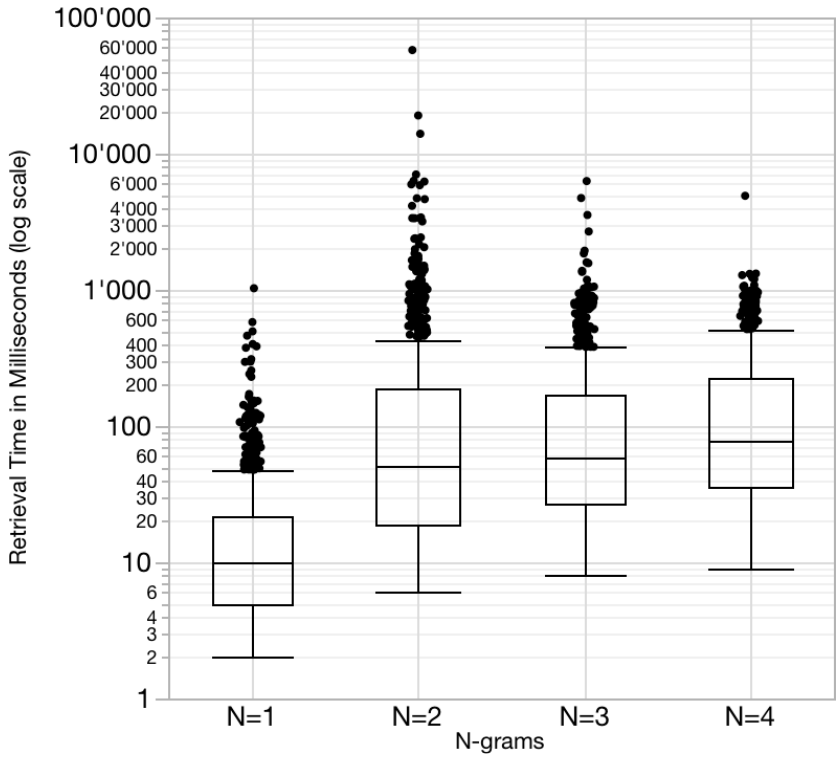
Performance

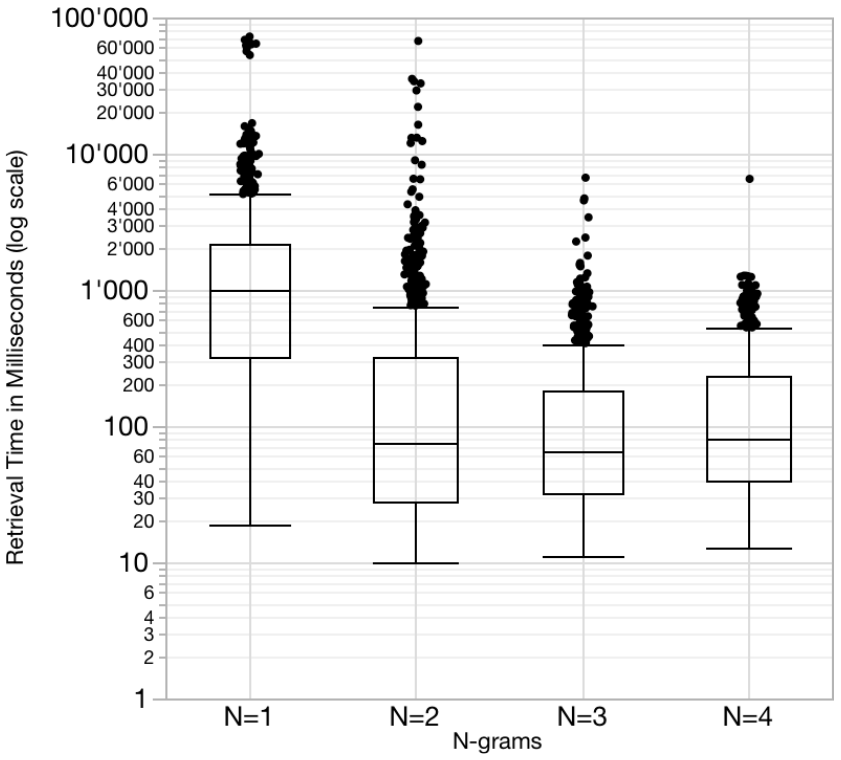
Hits and Translation Variants



Performance

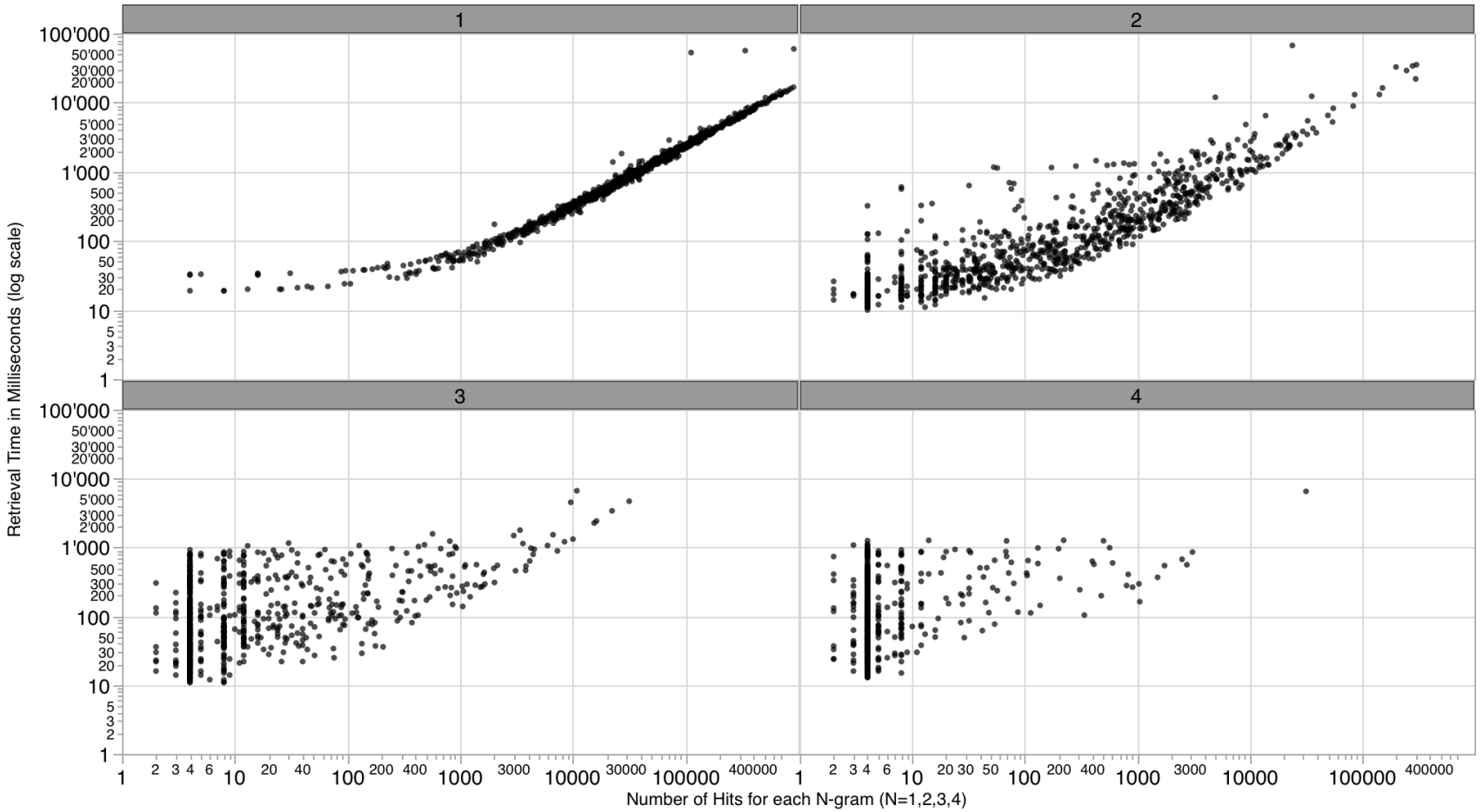
Hits and Translation Variants





Performance

Correlation of Number of Translation Variants and Retrieval Time



Outline

Conclusions

- We built an efficient corpus query system for exploration of multi-word units in large corpora.
  - Most multi-word queries (75 %) need less than 1 second.
- Materialized views provide an application-specific direct access to the required data.
- Database indexes allow for fast retrieval, but need to be adjusted to the particular use case (query).

Outlook

A new release of Multilingwis planned:

- two additional languages: Polish and Finnish
- sampling from search hits – 10 000 could be a reasonable limit
- export of retrieved data

Questions?

Tengo	una pregunta muy sencilla.
Ich	möchte eine sehr einfache Frage stellen.
I	have a very simple question.
Je	voudrais poser une question toute simple.
Ho	una domanda molto semplice.

Tengo	varias preguntas.
Ich	habe etliche Fragen.
I	have quite a few questions.
J'	ai quelques questions à poser.
Ho	varie domande.

En realidad	tengo algunas preguntas.
Ich	habe noch ein paar Fragen.
I am left with a few	questions.
J'	aurais encore quelques questions à poser.
Ho	ancora un paio di domande.

Tengo	una pregunta candente que hacer.
Ich	muss eine dringende Frage stellen.
I	have one burning question to ask.
J'	ai une question brûlante à poser.
Ho	una domanda urgente da sottoporre:

References I

Johannes Graën, Dolores Batinic, and Martin Volk (2014). "Cleaning the Europarl Corpus for Linguistic Applications". In: *Proceedings of the Conference on Natural Language Processing (KONVENS)*. (Hildesheim). Stiftung Universität Hildesheim

Martin Volk, Johannes Graën, and Elena Callegaro (2014). "Innovations in Parallel Corpus Search Tools". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*. (Reykjavik). European Language Resources Association (ELRA), pp. 3172–3178

References II

Johannes Graën and Simon Clematide (2015). "Challenges in the Alignment, Management and Exploitation of Large and Richly Annotated Multi-Parallel Corpora". In: *3rd Workshop on the Challenges in the Management of Large Corpora*. (Lancaster). Ed. by Piotr Bański et al. Institut für Deutsche Sprache, pp. 15–20

Simon Clematide, Johannes Graën, and Martin Volk (2016). "Multilingwis – A Multilingual Search Tool for Multi-Word Units in Multiparallel Corpora". In: *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives – Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües*. Ed. by Gloria Corpas Pastor. Geneva: Tradulex, pp. 447–455