Finding Translation Equivalents using Parallel Aligned Corpora

Johannes Graën

Institute of Computational Linguistics
University of Zurich

4th May, 2016

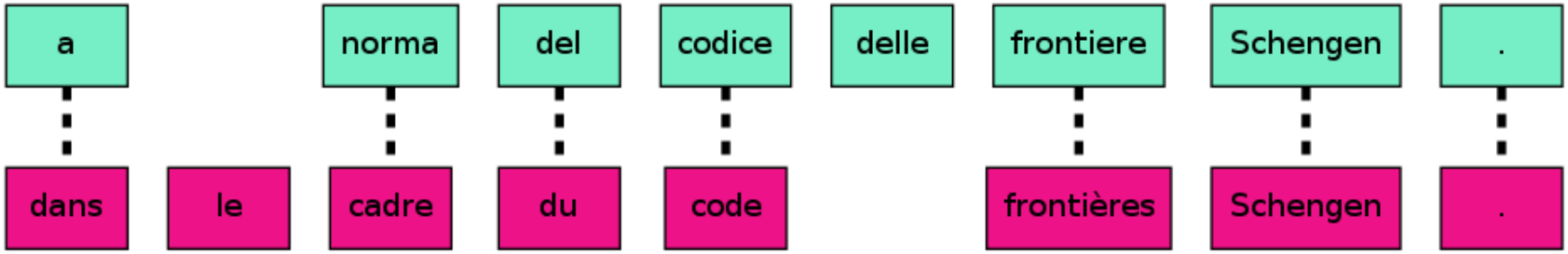
Outline

Outline

Project

SPARCLING

large-scale parallel corpora to study linguistic variation



Outline

Source

Europarl (version 7)

- comprises transcript of the European Parliament sittings
- contains numerous errors (Graën et al. 2014)
- has originally been compiled for training SMT systems
- features (reliable) alignment at the level of individual sittings

[1] http://pub.cl.uzh.ch/purl/costep

Source

Europarl (version 7)

- comprises transcript of the European Parliament sittings
- contains numerous errors (Graën et al. 2014)
- has originally been compiled for training SMT systems
- features (reliable) alignment at the level of individual sittings

CoStEP (Corrected & Structured Europarl Corpus; (ibid.))[1]

- bases on the Europarl Corpus files
- has undergone extensive cleaning
- comprehends ca. 75 % of the original corpus material
- features alignment of speaker turns and additional speaker information

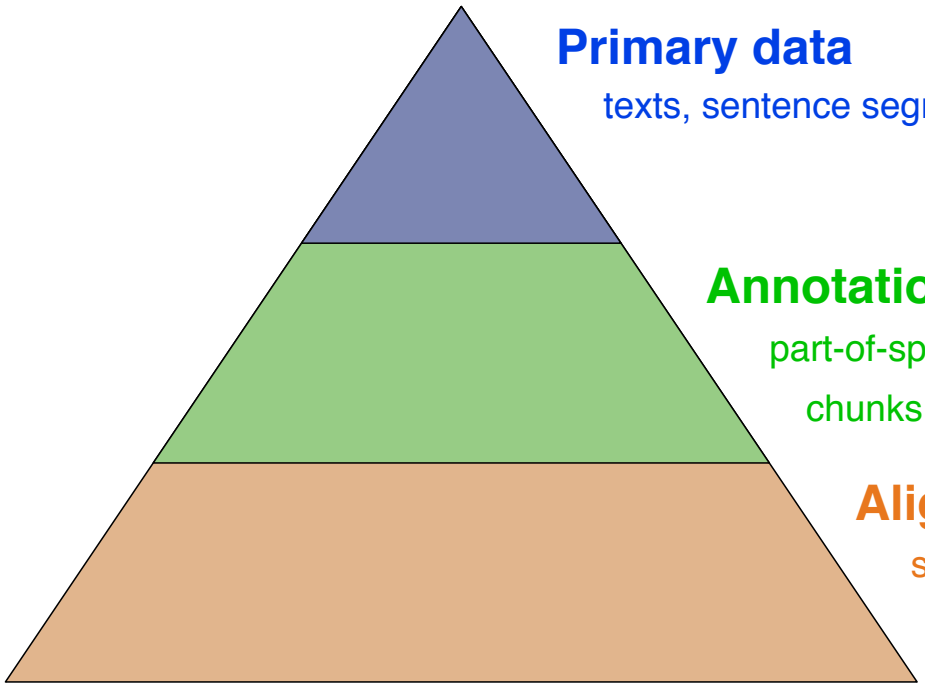
[1] http://pub.cl.uzh.ch/purl/costep

Our Corpus

Version 3

- 146.652 speaker turns from CoStEP in five languages: English, French, German, Italian and Spanish
- tokenization, part-of-speech tagging and lemmatization with the **TreeTagger** and its featured language models
- tag mapping to universal part-of-speech tags (uPoS)
- rule-based sentence segmentation
- pairwise sentence alignment with **hunalign**
- pairwise word alignment with **Giza++** based on lemmas of content words (ADJ, ADV, NOUN or VERB)
- when the lemmatizer did not assign a lemma to a particular token, we use the word form instead

Layout



Database-driven Corpus

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

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

Motivation

- we identified several types of online parallel corpus query systems (Volk et al. 2014)
- some 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

[2] https://glosbe.com/
[3] http://www.linguee.com/
[4] http://www.tradooit.com/

Linguee


Linguee | Dictionary for
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
ä ö ü ß
English ↔ German
English → German
German → English
English ↔ Portuguese
English ↔ Spanish
English ↔ French
English ↔ Italian
English ↔ Russian
English ↔ Japanese


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

Concept

Multilingwis

Multilingual word information system

We designed Multilingwis to be a corpus exploration system that

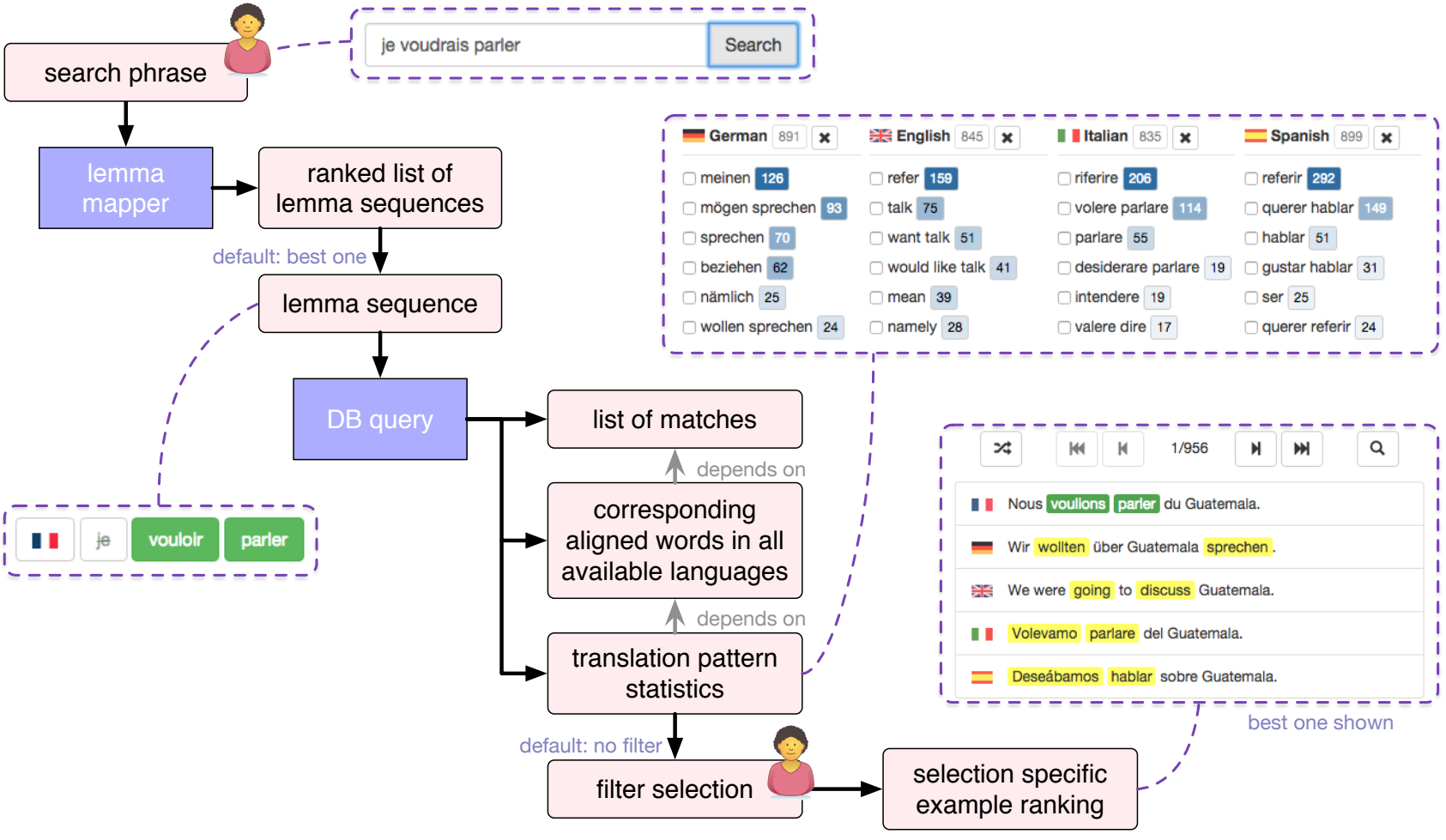
- allows for similar ad-hoc searches
- shows the distribution of translation variants
- offers a backward search for each of those variants
- provides examples with translation equivalents marked

Another important aspect: Speed!

Procedure

- 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

Procedure


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

Demo

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

Indexes

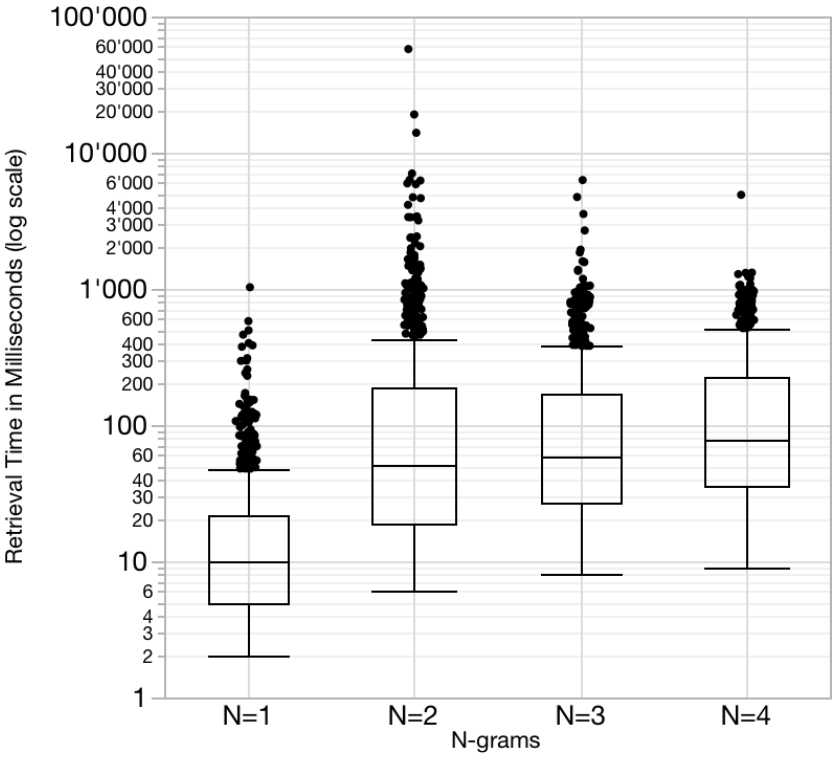
- materialized view on lemmas and relevant foreign keys
- composite index over all columns starting with the lemma (7.3 GB for 220 million rows)
- another composite index on symmetrized view of word alignments (null alignments skipped);
  we use the 'union' symmetrization method for better recall
  (9.0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

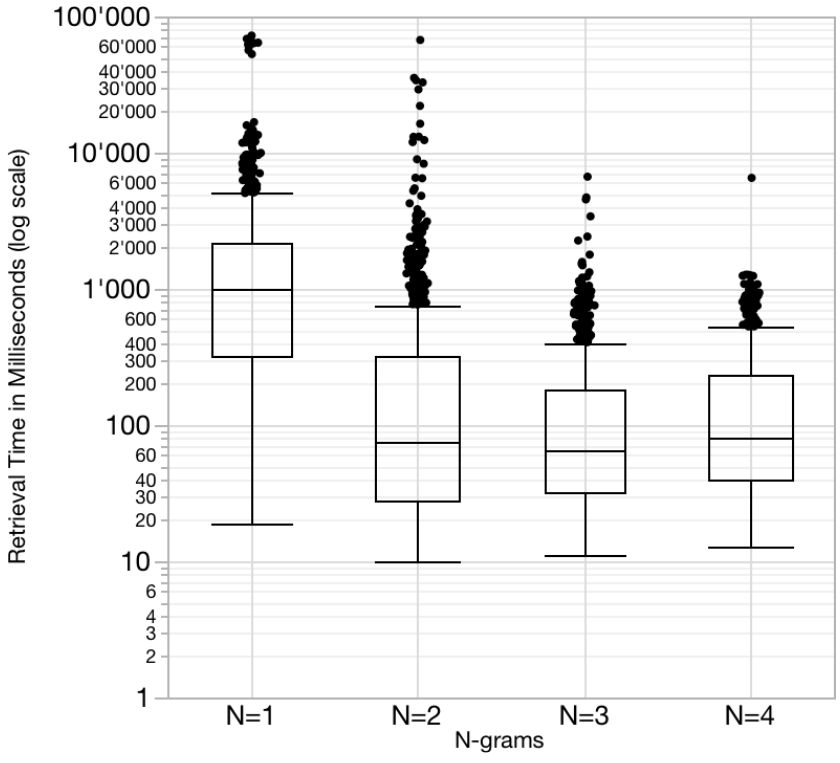
Query

- the lemma index is scanned to identify all hits
- translation equivalents are retrieved by intersecting hits with the alignment index
- lemmas of all aligned tokens are joined and frequencies of lemma sequences, i.e. translation variants, are aggregated
- a particular search function is responsible for each count of source lemmas, allowing for pre-planned queries

Performance

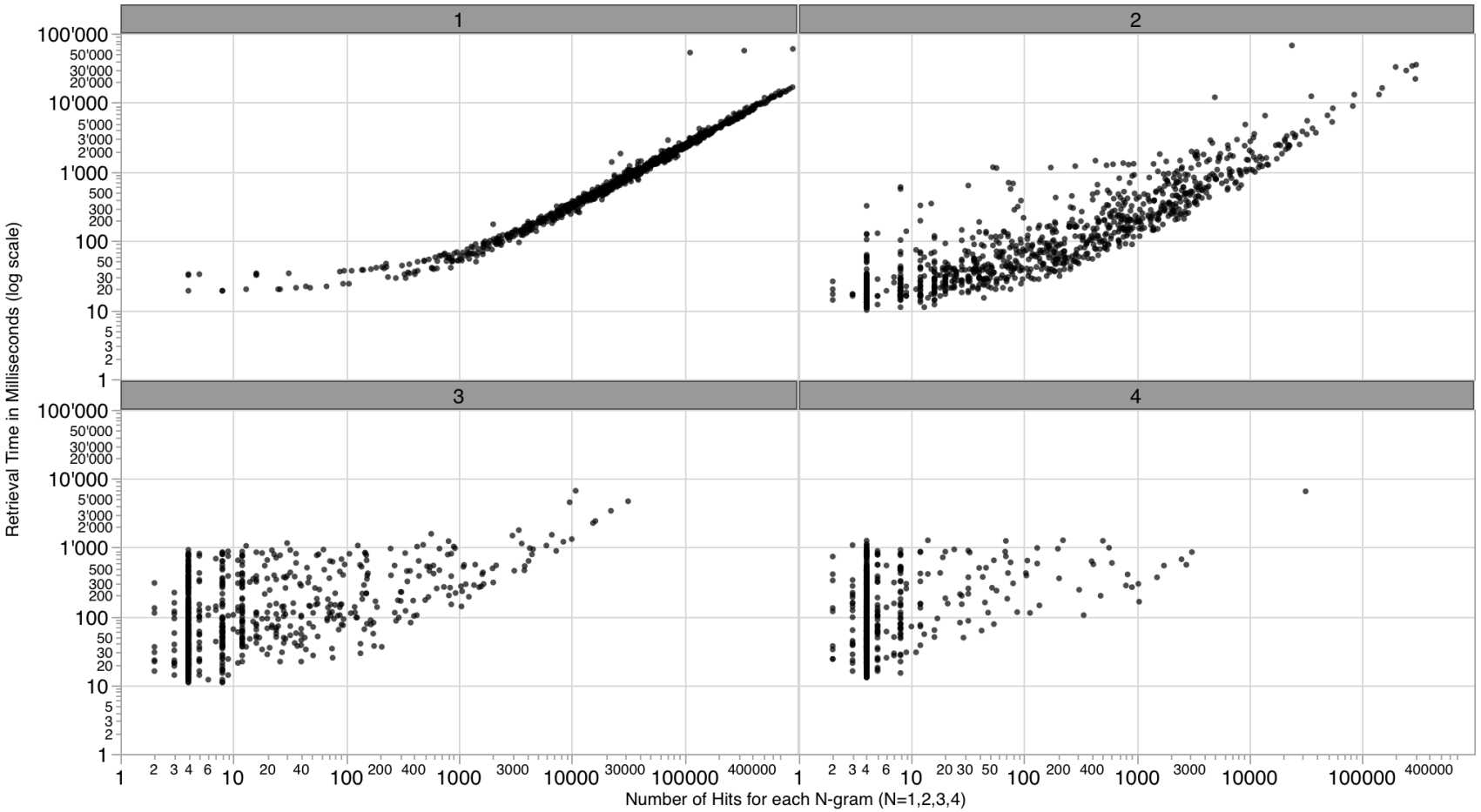
Hits (left) and Translation Equivalents (right)





Performance

Correlation of Number of Translation Equivalents and Retrieval Time



Outline

More Complex Queries

Reattaching Separable Verb Prefixes in German

Example

- 70 % der tschechischen Bürger lehnen das System ab.
- Die Schweiz stellt ein viel ernsteres Problem dar.
- Wer die Verhandlung verlässt, stimmt zu.
- Daraufhin schlug er die Unabhängigkeit des Kosovo vor.

More Complex Queries

Reattaching Separable Verb Prefixes in German

- find separated verb prefix (PTKVZ)
- seek the next finite verb (VVFIN or VVIMP) to the left
- check whether the recombined verb exists in the corpus
- retrieve statistics on the translation equivalents of the verbs

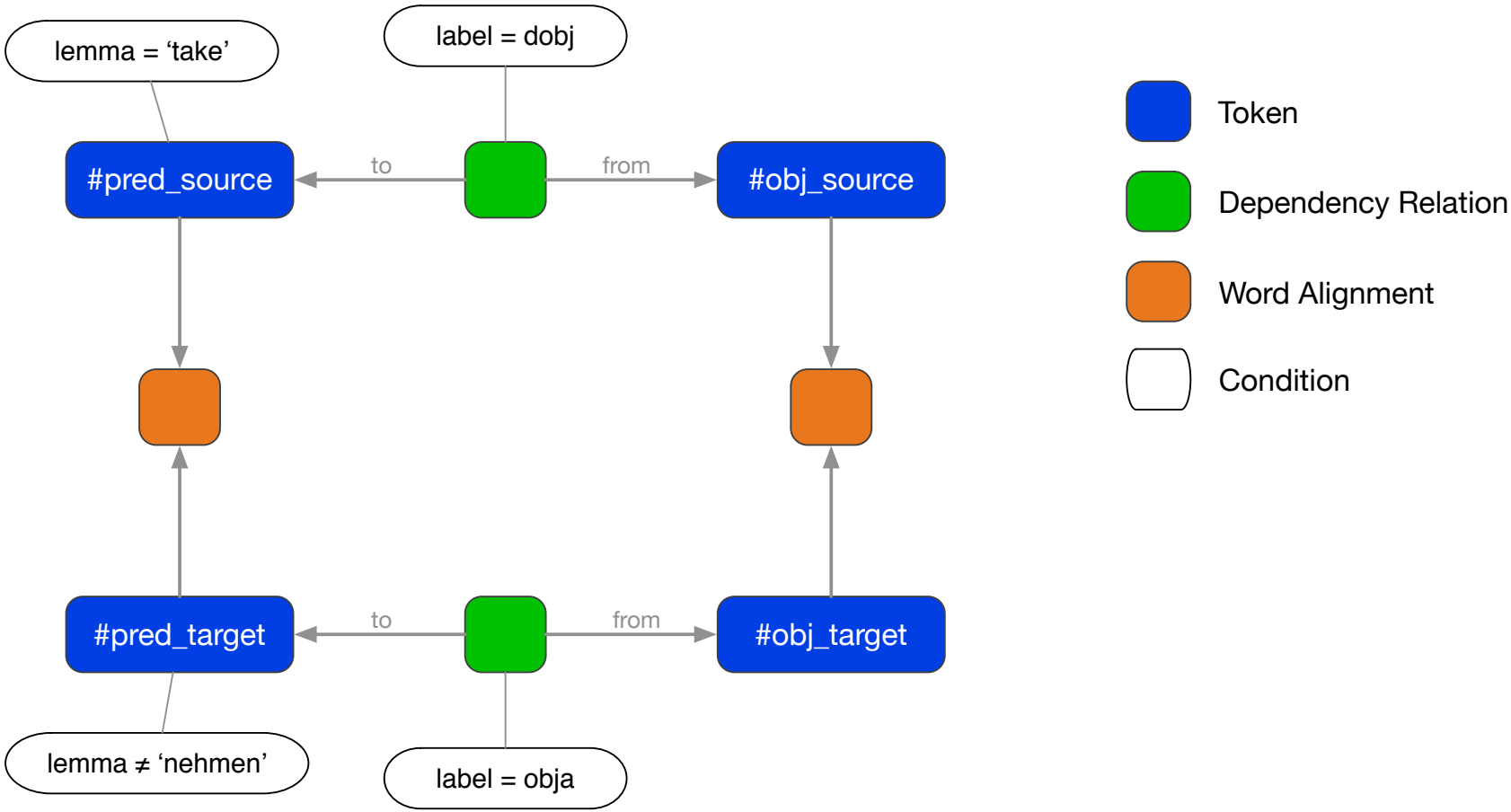
More Complex Queries

Reattaching Separable Verb Prefixes in German

prefix	**verb**	**en**	**it**
dar	stellen	be (59.2 %) represent (15.0 %)	rappresentare (30.0 %) essere (26.3 %) costituire (23.0 %)
zu	stimmen	agree (80.3 %)	concordare (51.7 %) condividere (15.8 %)
vor	schlagen	propose (76.4 %) suggest (17.6 %)	proporre (85.2 %)
ab	hängen	depend (86.4 %)	dipendere (89.0 %)

More Complex Queries

Finding Light Verb Constructions



More Complex Queries

Finding Light Verb Constructions

Example

en	**de**
take decision	Entscheidung treffen
	'strike/hit/score a decision'
take opportunity	Gelegenheit ergreifen
	'seize/pounce/grasp an opportunity'
take sanction	Sanktion verhängen
	'impose/declare/veil a sanction'
⋮	⋮

General Approach

- describe corpus query pattern in terms of relations and conditions
- retrieve all hits and calculate cooccurrence frequencies for the source language
- retrieve translation equivalents for all hits and calculate cooccurrence frequencies for the target languages
- calculate cooccurrence frequencies for the translation relations based on all alignments

Demo

https://pub.cl.uzh.ch/projects/sparcling/yqdemo/

Outline

Outlook

- Multilingwis

Outlook

- Multilingwis
  - efficient tool for exploration of translation variants

Outlook

- Multilingwis
  - efficient tool for exploration of translation variants
  - upcoming user evaluation

Outlook

- Multilingwis
  - efficient tool for exploration of translation variants
  - upcoming user evaluation
  - new release planned (7 languages, better alignments, bug-fixes)

Outlook

- Multilingwis
  - efficient tool for exploration of translation variants
  - upcoming user evaluation
  - new release planned (7 languages, better alignments, bug-fixes)
- Ongoing work

Outlook

- Multilingwis
  - efficient tool for exploration of translation variants
  - upcoming user evaluation
  - new release planned (7 languages, better alignments, bug-fixes)
- Ongoing work
  - focus on inter-lingual statistical measures

Outlook

- Multilingwis
  - efficient tool for exploration of translation variants
  - upcoming user evaluation
  - new release planned (7 languages, better alignments, bug-fixes)
- Ongoing work
  - focus on inter-lingual statistical measures
  - allow for the user to define the initial corpus query pattern

Tengo preguntas que hacerles.	Tengo una pregunta complementaria.	Solamente tengo una pregunta adicional.
Ich möchte Ihnen einige Fragen stellen.	Ich habe noch eine Zusatzfrage.	Ich habe nur noch eine andere Frage.
I have questions for you.	I have a supplementary question.	I have just one other query.
J'ai une question pour vous.	J'ai une question complémentaire à formuler.	J'aurais une autre question.
Avrei delle domande da porvi.	Avrei un'interrogazione complementare.	Vorrei porre un altro quesito.
Tengo una pregunta candente que hacer.	Tengo una pregunta a este respecto.	No obstante, tengo una pregunta:
Ich muss eine dringende Frage stellen.	Dazu habe ich allerdings eine Frage.	Ich habe indes eine Frage:
I have one burning question to ask.	I have a question on this point.	I have, however, a question:
J'ai une question brûlante à poser.	J'ai cependant une question à ce sujet.	J'ai, cependant, une question :
Ho una domanda urgente da sottoporre:	Ho una domanda al riguardo.	Mi sia permessa una domanda:
No obstante, tengo una pregunta:	Tengo una pregunta muy sencilla.	Tengo una pregunta sobre el calendario.
Ich habe jedoch eine Frage:	Ich möchte eine sehr einfache Frage stellen.	Mich beschäftigt eine Frage zum Timing.
But I also have a question:	I have a very simple question.	I have a question on the timetable.
J'ai tout de même une question :	Je voudrais poser une question toute simple.	J'ai une question sur le calendrier.
Avrei però una domanda:	Ho una domanda molto semplice.	Ho una domanda sul calendario.
Tengo varias preguntas.	En realidad tengo algunas preguntas.	Para terminar, tengo dos preguntas.
Ich habe etliche Fragen.	Ich habe noch ein paar Fragen.	Abschließend noch zwei Fragen:
I have quite a few questions.	I am left with a few questions.	To finish, I have two questions.
J'ai quelques questions à poser.	J'aurais encore quelques questions à poser.	Je terminerai par deux questions.
Ho varie domande.	Ho ancora un paio di domande.	Per concludere, ho due domande da porre.

References I

Martin Volk et al. (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

Johannes Graën et al. (2014). "Cleaning the Europarl Corpus for Linguistic Applications". In: *Proceedings of the Conference on Natural Language Processing (KONVENS)*. (Hildesheim). Stiftung Universität Hildesheim

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 et al. (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

References III

Johannes Graën et al. (2016). "Efficient Exploration of Translation Variants in Large Multiparallel Corpora Using a Relational Database". In: forthcoming