



University of  
Zurich<sup>UZH</sup>

Institute of Computational Linguistics

---

multiling*wis*

# A Multilingual Search Tool for Multi-word Units in Multiparallel Corpora

Johannes Graën, Simon Clematide, Martin Volk — 29.06.2015

<http://pub.cl.uzh.ch/purl/multilingwis>

# Outline

- Motivation
- Existing search tools
- Data preparation steps
- Using multilingwis
- Limitations / Future Work
- Conclusions

The screenshot displays the 'multilingwis' web application interface. The browser address bar shows the URL: <https://pub.cl.uzh.ch/projects/sparcling/multilingwis/?l=overview...>. The page title is 'multilingwis europarl edition 1.0 [ en es de fr it ]'. A search bar contains the text 'overview' and a 'Search' button. Below the search bar, there are language selection buttons for German, French, Italian, and Spanish. The main content area shows a list of search results for the term 'overview' in various languages, with counts for each result. The results are:

Language	Term	Count
German	Überblick	61
German	Übersicht	17
French	aperçu	28
French	vue ensemble	24
French	vision	6
Italian	panoramica	35
Italian	quadro	12
Italian	visione insieme	10
Spanish	visión general	19
Spanish	visión	18
Spanish	visión conjunto	10
Spanish	resumen	8
Spanish	visión global	6

Below the list, there are navigation controls including a search icon, a refresh icon, a page number '15/140', and a search icon. The main content area displays the first result in English: 'It gives an impressive overview of the situation in Ukraine at a crucial juncture in the country's history.' Below this, there are translations in German, French, Italian, and Spanish, each with the corresponding term highlighted in yellow.

German: Er vermittelt einen eindrucksvollen Überblick über die Situation in der Ukraine, die sich derzeit an einem entscheidenden Wendepunkt in ihrer Geschichte befindet.

French: Ce rapport donne un aperçu saisissant de la situation de l'Ukraine à un moment crucial de son histoire.

Italian: Il documento fornisce una straordinaria visione d'insieme della situazione in Ucraina in un momento cruciale della storia del paese.

Spanish: Ofrece una impresionante visión general de la situación en Ucrania en una coyuntura crucial de su historia.

At the bottom of the page, there is a footer: 'Institute of Computational Linguistics | University of Zurich | Realized by J. Grahn / S. Ciernatide as part of the SPARCLING project.'



# Motivation

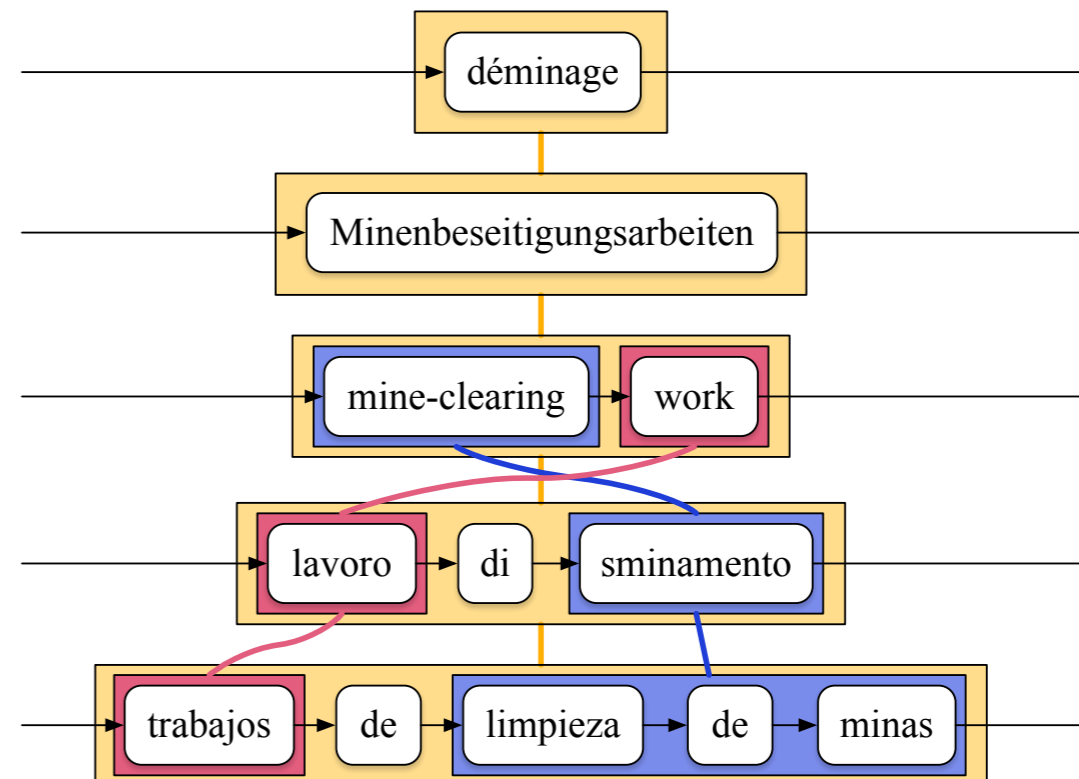
# Motivation

## User needs

- Language learners
  - Typical translations / translation variants
  - Usage contexts
- Translators
  - Translation variants with frequencies (probably genre and domain specific)
- (Corpus) linguists
  - Full-fledged corpus query tools

# Motivation SPARCLING

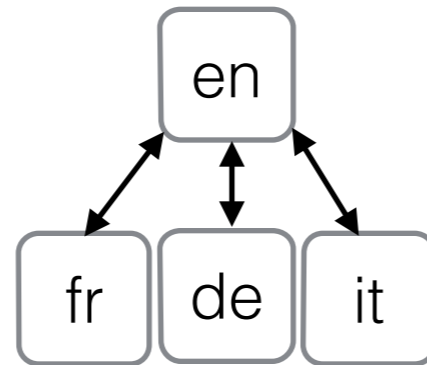
- Tiger/TreeAligner query language adapted for querying multiparallel data with several layers of linguistic data
- Inter-lingua (tokenization, part-of-speech tagging, chunking, dependency parsing, coreference resolution, ...)
- Intra-lingua (alignments on text, sentence, word and sub-sentential level)



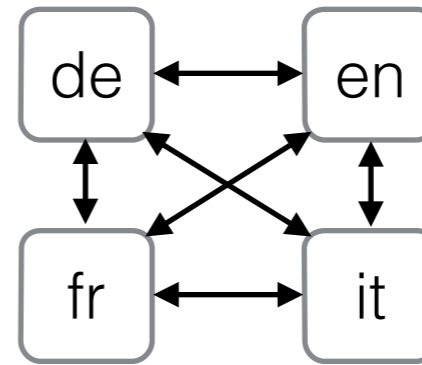
# Motivation

## Our goal for multilingwis

- Build an empirical multiparallel tool for translation spotting (including MWUs)



Multilingual parallel  
(*Medline titles*)



Multiparallel  
(*Europarl*)

- Provide a user-friendly search interface (addressing also non-linguists) for ad-hoc searches
- Facilitate the user to explore translation variants





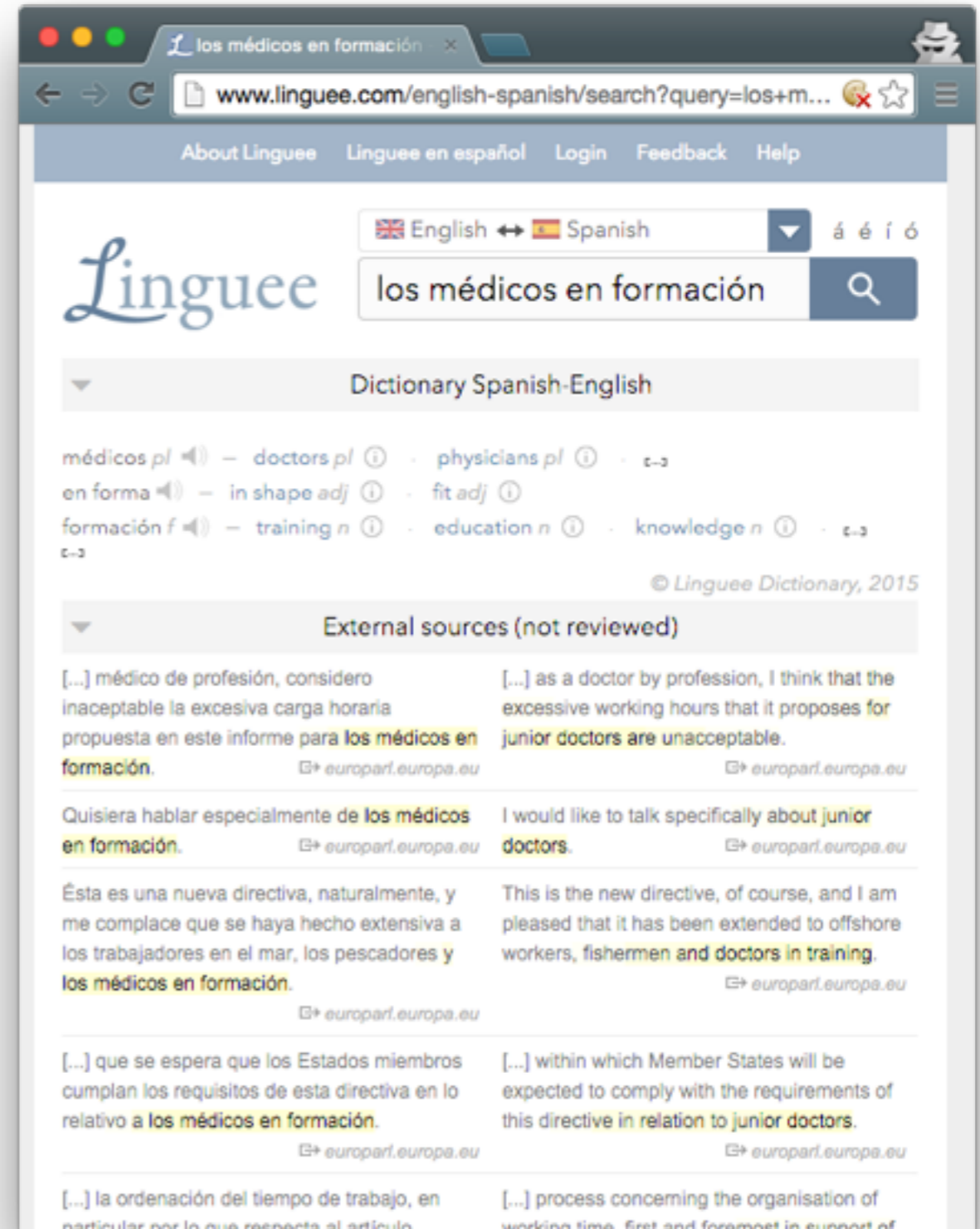
# Existing search tools

# Existing search tools

- Online dictionaries with examples
- Translation search tools
- Linguistic corpus query engines
- Other concordancing tools

# Existing search tools

## Linguee



# Existing search tools


## Tradooit



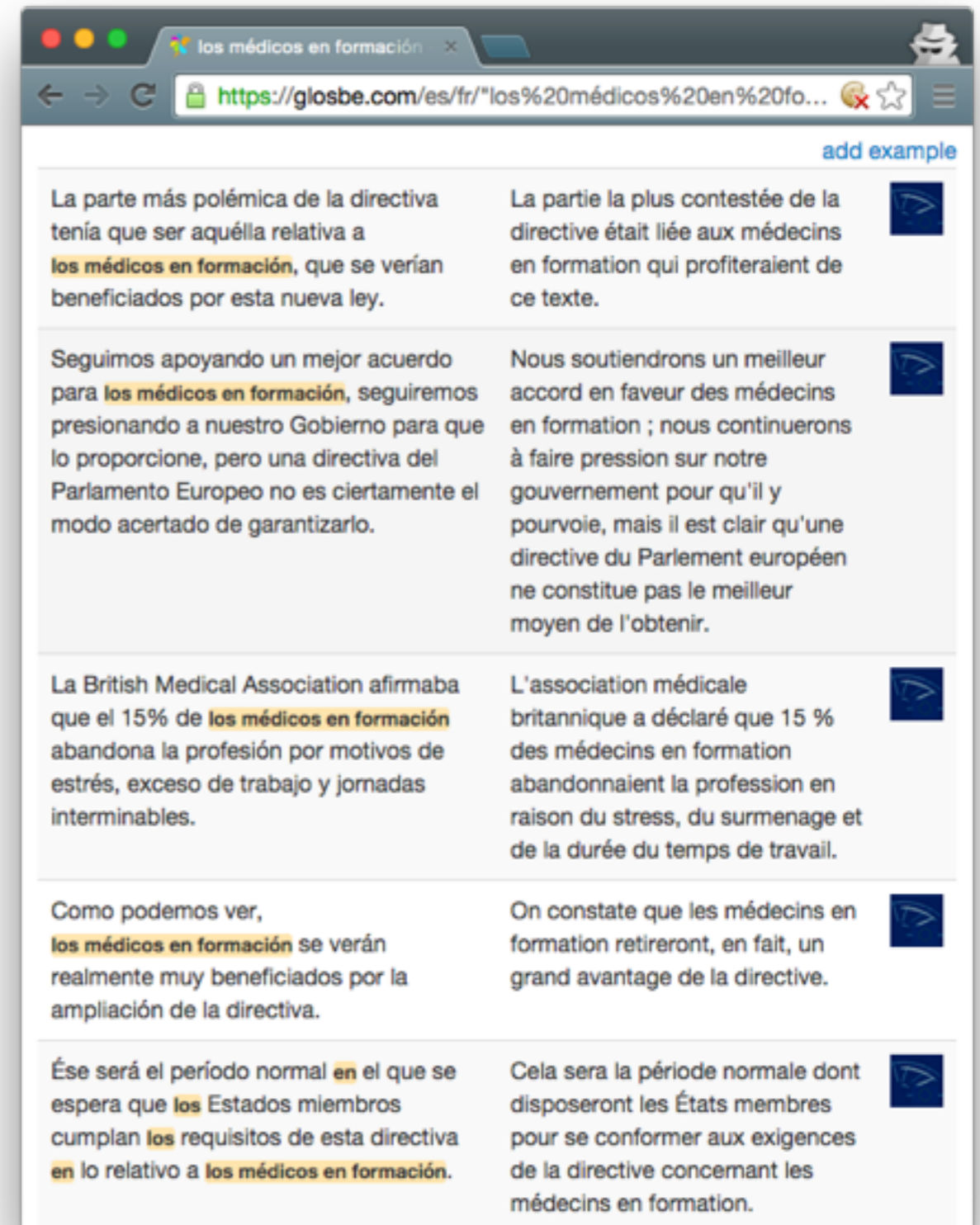


# Existing search tools

## Glosbe



The screenshot shows the Glosbe website homepage. At the top, there is a navigation bar with the Glosbe logo and a hamburger menu. Below this is a large heading: "Glosbe - the multilingual online dictionary". A search bar is prominently displayed, labeled "Search German - English Dictionary". Below the search bar, there are dropdown menus for "German" and "English", a double-headed arrow icon, and buttons for "ca", "de", "en", and "es". A paragraph of text states: "We provide free dictionaries for almost every existing language and translation memory with 1 013 284 995 sentences included." Below this, there is a bulleted list of features: "Almost every live language.", "Huge dictionary database.", "Millions of examples.", and "Unique phrases and expressions." At the bottom, there is a link: "If your language is not listed in select boxes try all dictionaries link."



The screenshot shows a search result page on Glosbe. The browser address bar shows the URL: "https://glosbe.com/es/fr/"los%20médicos%20en%20fo...". The page displays a grid of example sentences in two columns. Each row contains a Spanish sentence on the left and its French translation on the right. The Spanish sentences are: "La parte más polémica de la directiva tenía que ser aquella relativa a los médicos en formación, que se verían beneficiados por esta nueva ley.", "Seguimos apoyando un mejor acuerdo para los médicos en formación, seguiremos presionando a nuestro Gobierno para que lo proporcione, pero una directiva del Parlamento Europeo no es ciertamente el modo acertado de garantizarlo.", "La British Medical Association afirmaba que el 15% de los médicos en formación abandona la profesión por motivos de estrés, exceso de trabajo y jornadas interminables.", "Como podemos ver, los médicos en formación se verán realmente muy beneficiados por la ampliación de la directiva.", and "Ése será el período normal en el que se espera que los Estados miembros cumplan los requisitos de esta directiva en lo relativo a los médicos en formación." The French sentences are: "La partie la plus contestée de la directive était liée aux médecins en formation qui profiteraient de ce texte.", "Nous soutiendrons un meilleur accord en faveur des médecins en formation ; nous continuerons à faire pression sur notre gouvernement pour qu'il y pourvoie, mais il est clair qu'une directive du Parlement européen ne constitue pas le meilleur moyen de l'obtenir.", "L'association médicale britannique a déclaré que 15 % des médecins en formation abandonnaient la profession en raison du stress, du surmenage et de la durée du temps de travail.", "On constate que les médecins en formation retireront, en fait, un grand avantage de la directive.", and "Cela sera la période normale dont disposeront les États membres pour se conformer aux exigences de la directive concernant les médecins en formation." Each row also includes a small blue icon with a white checkmark.

# Existing search tools

## TAUS Data

The screenshot shows the TAUS Data website interface. At the top, the browser address bar displays <https://www.tausdata.org/index.php/data>. The TAUS logo is prominently displayed, with the tagline "Human Language Project". A navigation menu includes links for HOME, ABOUT US, MEMBERSHIP, TECHNOLOGY, and DATA. A "BUY DATA" button is also visible.

### LANGUAGE DATA

Total number of language pairs in repository: **2214** | Total number of words in repository: **59,551,108,437**

Language selection: English (Canada) > French (France)

Include Matrix TM Results (?)

Industry: Any...

Available words (?): **6,854**

- Direct TM words (?): **0**
- Matrix TM words (?): **6,854**

You need credits to download TMs. To get credits, you can:

- [Upload TMs](#)
- [Buy credits](#)

# Existing search tools

## Bwananet

**4. Consulta: concordança estàndard multilingüe**

Selecció feta:  
Llengua dels documents: Català  
Documents paral·lels: Castellà - Anglès  
Documents seleccionats: Tot el corpus  
quantitat de documents: 107  
Nombre de paraules: 1580569

**a) Informació específica sobre la concordança** [Afegir més columnes](#)

Unitats	<>	Unitat #1	Unitat #2	Unitat #3	Unitat #4	Unitat #5	</>
- Formes							
- Lemes							
- Categories	<input type="checkbox"/>	-	-	-	-	-	<input type="checkbox"/>
Repetició		-	-	-	-	-	
Negació		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Ordenat per	<input checked="" type="radio"/> no	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
		Ordre alfabètic per: <input type="radio"/> Formes <input checked="" type="radio"/> Lemes					

**b) Altres informacions necessàries**

Context:  Complet  Parcial +/- 5 (unitats a dreta i esquerra)

Presentació de la concordança:  Formes  Lemes  Categories

Informació addicional:  Estatus del document  Subdomini  Tipus de document

Quantitat de resultats:  primers resultats

[Restriccions en llengües d'arribada](#)

Buscar Cancel·lar la selecció Ajuda

# Existing search tools bilingwis

Universität Zürich search page | about **bilingwis** | language: DE EN | help

## bilingwis

finding translations in bilingual context

Language pair: DE <> FR

corpus: SAC year books 1957-2013

search direction: DE > DE+FR

search by: lemma

sort results by: frequency

case-sensitive search

ä ö ü ß Ä Ö Ü

bilingwis ver. 3.72  
 © University of Zurich  
 Institute of Computational Linguistics  
 last update 11.08.2014

Universität Zürich search page | about **bilingwis** | language: DE EN | help

montée  
— 123 hits  
(100 shown)

1957 S. Walcher: <i>Bergfahrten im Zillertal</i> Orig: DE	Links von ihr (im Sinne des <b>Anstieges</b> ) ist der steile Fels gut gegliedert.	A sa gauche (dans le sens de la <b>montée</b> ) le rocher raide offre une excellente varappe.
	den <b>Anstieg</b> zum Frankbachjoch kannten wir, und das Wetter war schön.	Nous n'avons pas de raison de nous hâter, puisque nous connaissons la <b>montée</b> du Col de Frankbach et que le temps reste beau.
	Der <b>Anstieg</b> zu unseren versteckten Säcken war nicht gerade angenehm, konnte aber die Freude über die gelungene Giga-litzüberschreitung nicht trüben.	La <b>montée</b> vers le dépôt de nos sacs manque de charme.

ascension  
— 30 hits

1957 S. Walcher:	Wir mussten dabei die Erfahrung machen, dass der letzte Teil des <b>Anstieges</b> , der Übergang vom Vor-zum Hauptgipfel durch einen	Nous y faisons la découverte que la dernière partie de l' <b>ascension</b> , exactement la traversée de l'avant-sommet ou sommet principal, est
---------------------	--	---





# Data Preparation

# Data Preparation

## Language-wise

- Extraction of parallel texts in English, French, German, Italian and Spanish from the *Corrected & Structured Europarl Corpus (CoStEP)*
- Tokenization and tagging with the *TreeTagger* (adapted)
  - Mapping of the particular tagsets to *universal part-of-speech tags* (12 different tags)
- Rule-based sentence segmentation

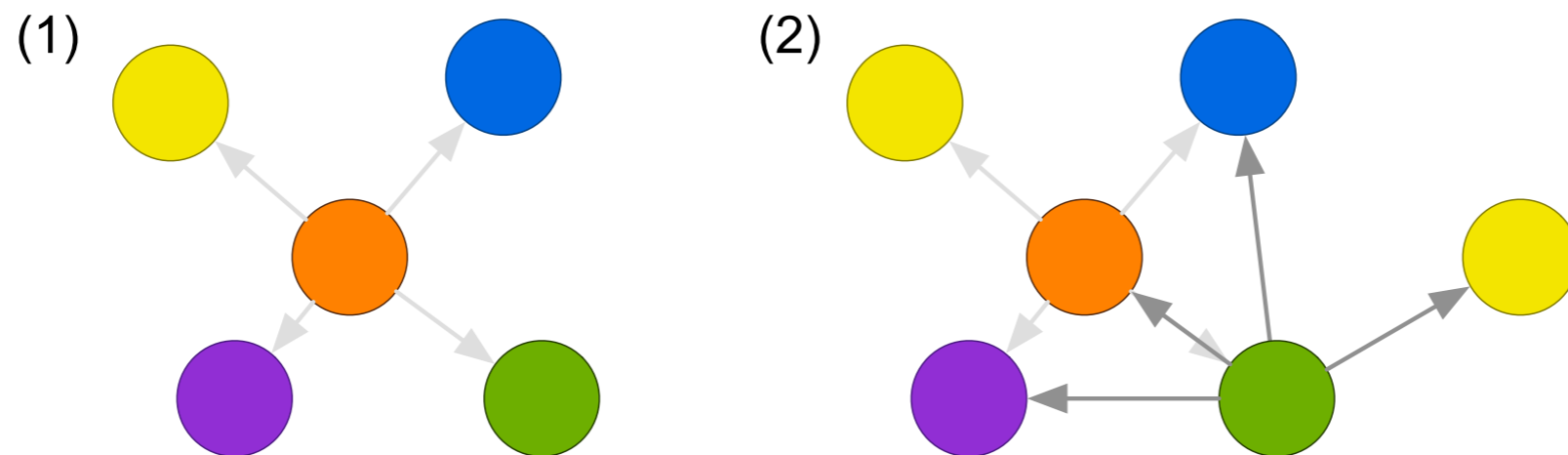
# Data Preparation

## Alignments

- Bilingual sentence segment alignment with *hunalign* for each language pair
  - Based on lemmas identified by the *TreeTagger*
- Bilingual word alignment with *Giza++* for each language pair and both directions
  - Based on sequences of content words (i.e. those being tagged as either NOUN, VERB, ADJ or ADV)

# Data Preparation Features/Challenges

- Fully automated (large amounts of data)
- All steps show certain error rates
- ... which accumulate
- No multiparallel sentence and word alignment



# Data Preparation Statistics — Tokens, Words and Lemmas

<b>Texts</b>	733'260
<b>Segments</b>	8'471'061
<b>Tokens</b>	219'523'637
<b>Words</b>	792'242
<b>Lemmas</b>	214'585

# Data Preparation Statistics — Lemma distribution

Lemmas	214'585
English	43'993
French	27'737
German	97'311
Italian	29'058
Spanish	16'486

# Data Preparation

## Statistics — Content word distribution

<b>110'425'468</b>	<b>NOUN</b>	<b>VERB</b>	<b>ADJ</b>	<b>ADV</b>
<b>English</b>	10'396'359	7'100'017	3'182'057	2'267'892
<b>French</b>	10'507'497	6'427'094	3'166'465	2'598'792
<b>German</b>	9'108'400	5'691'497	3'427'933	2'167'203
<b>Italian</b>	10'431'079	6'266'542	4'013'653	2'176'171
<b>Spanish</b>	10'211'886	6'435'966	3'338'786	1'510'179



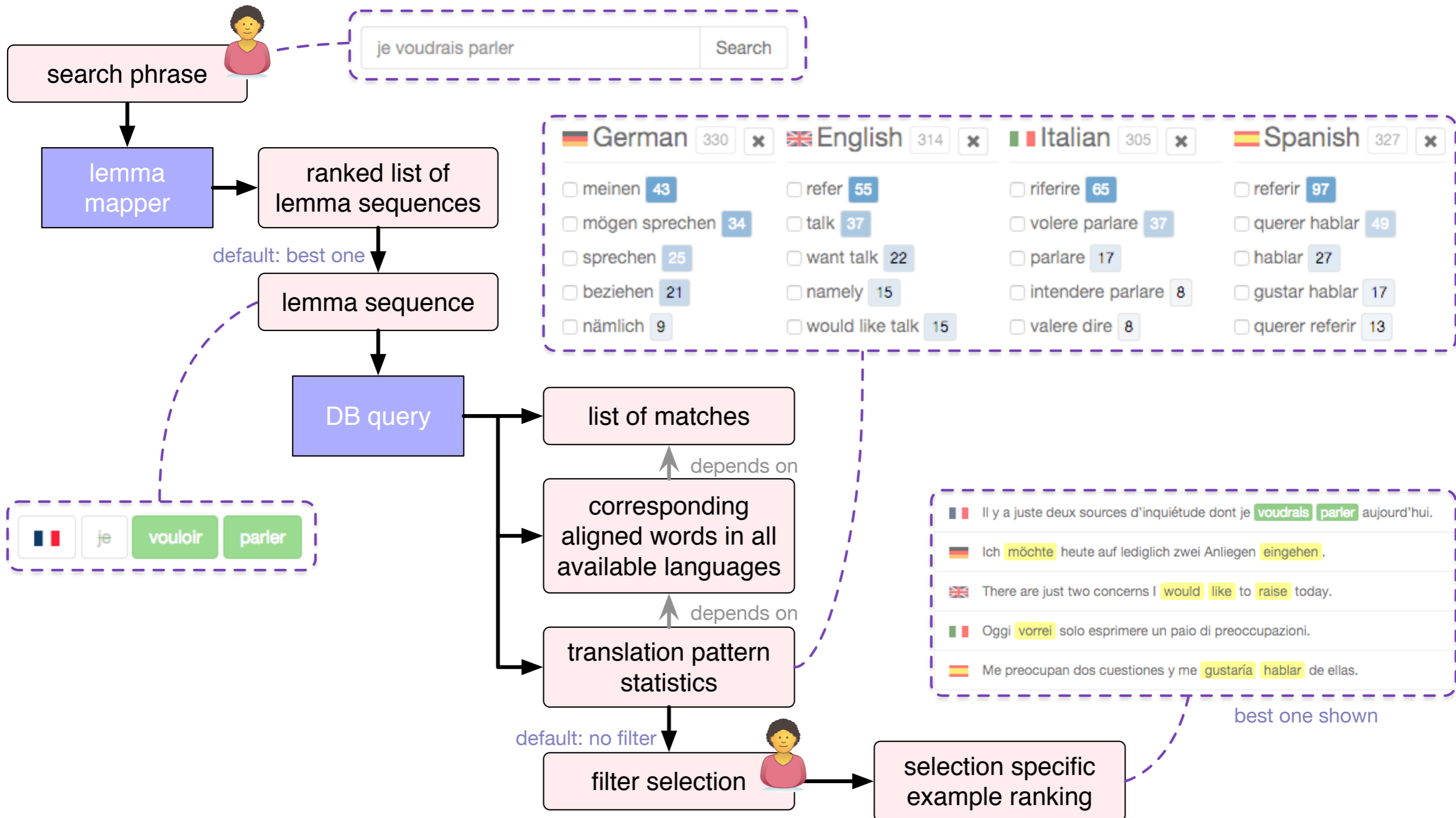


# Using multilingwis

# Using multilingwis Workflow

- User input gets lemmatized, function words are removed
- Sequence of lemmas is searched in a database
  - A maximum amount of 3 non-content words between the search lemmas is allowed
- For each hit, the aligned words are searched for and a frequency distribution of translation variants is calculated thereupon
- The overall best (=shortest) example is shown
- Every translation variant can again be queried for with a single click

# Using multilingwis Workflow



# Using multilingwis

## Restricting the search space

The screenshot shows a search interface with four language-specific search filters and their corresponding results:

- German:** Filter: 🇩🇪 Menschenrechtsverletzung
- English:** Filter: 🇬🇧 ✓ violation human right 131
- Italian:** Filter: 🇮🇹 ✓ violazione diritto uomo 101
- French:** Filter: 🇫🇷 ✓ violation droit homme 401

The search results are displayed in a list format with navigation controls (back, forward, search, etc.). The results are:

- German:** Was Äthiopien betrifft, so finde ich seit zwei Jahren kein Gehör, wenn es um **Menschenrechtsverletzungen** geht.
- English:** For two years, people have been knocking on my door about the **violation** of **human rights**.
- French:** Depuis deux ans, on s'adresse à moi pour des questions de **violation** des **droits** de l' **homme**.
- Italian:** Da due anni mi vengono comunicate **violazioni** dei **dritti** dell' **uomo**.

# Using multilingwis

## Behind the scenes

- Lemmatization and language identification via finite state transducer trained on the data available (word, lemmas, part-of-speech tags, frequencies)
  - Mapping of wrong cases, character variants, ...
- Relational database searches for occurrences of a given lemma sequence, intersects the result with the word alignments stored and aggregates the translation variants
- The user interface shows the results of both steps and allows the user to restrict the alignment search space by selecting one or more translation variants



# Limitations / Future Work

- Data preparation
  - Lemma-based word alignment could be improved by harmonizing Lemmas across languages (e.g. removing gender suffix in German)
- User interface
  - Add part-of-speech filtering (“el cotejo\_N de ...”)
  - Option to choose the 2nd, 3rd, ... best lemmatization
- Export of the query results
- User testing / feedback



# Conclusions


# Conclusions

- Efficient search tool for multiparallel corpora that supports
  - searches for multi-word units in
  - any available language
- Simple user interface for
  - ad-hoc searches and
  - exploration of translation variants

<http://pub.cl.uzh.ch/purl/multilingwis>

have question Questions? Search

 have question

 **German** 148 ✕ ☰


- haben Frage 64
- Frage 50
- Anfrage 7

 **French** 146 ✕ ☰

- avoir question 41
- question 32
- poser question 22
- avoir question poser 12

 **Italian** 146 ✕ ☰

- avere domanda 49
- domanda 15
- interrogazione 12
- volere domanda 11
- desiderare domanda 8

 **Spanish** 148 ✕ ☰

- tener pregunta 66
- pregunta 27
- querer pregunta 16

↻
⏮
⏪
13/149
⏩
⏭
🔍

-  I have questions for you.
-  Ich möchte Ihnen einige Fragen stellen.
-  J' ai une question pour vous.
-  Avrei delle domande da porvi.
-  Tengo preguntas que hacerles.

