Korpuslinguistik mit PostgreSQL

Johannes Graën

Institut für Computerlinguistik
Universität Zürich
graen@cl.uzh.ch

2016-06-24

Übersicht

Übersicht

Korpuslinguistik

Was ist das?

Korpuslinguistik

Was ist das?

der Korpus

Korpuslinguistik

Was ist das?

das Korpus

- digitale Textsammlung
- strukturiert
- abfragbar
- liefert empirische Daten für die Sprachforschung

Korpuslinguistik

Was ist das?

parallele Korpora

- digitale Textsammlung sich einander entsprechender Texte in mehreren Sprachen
- mehrheitlich Übersetzungen (bei einer Sprachversion handelt es sich i.d.R. um das Original)
- Dimensionen u.a.: Anzahl Token (Wörter, Satzzeichen etc.), Anzahl der Sprachen, Annotationsebenen (zusätzliche Informationen)

Unser Korpus I

basierend auf 15 Jahren Debatten des Europäischen Parlaments

- mehr als 140.000 Redebeiträge (entspricht 220 Mio. Token)
- in fünf Sprachen: **de**, **en**, **es**, **fr**, **it**
- Wortarten-Erkennung (*part-of-speech tagging*)
- Lemmatisierung (Reduzierung der flektierten Formen auf eine Grundform, z.B. Infinitiv)
- Satzalignierung (Verknüpfung sich entsprechender Sätze)
- Wortalignierung (Verknüpfung sich entsprechender Token)
- syntaktische Dependenzrelationen

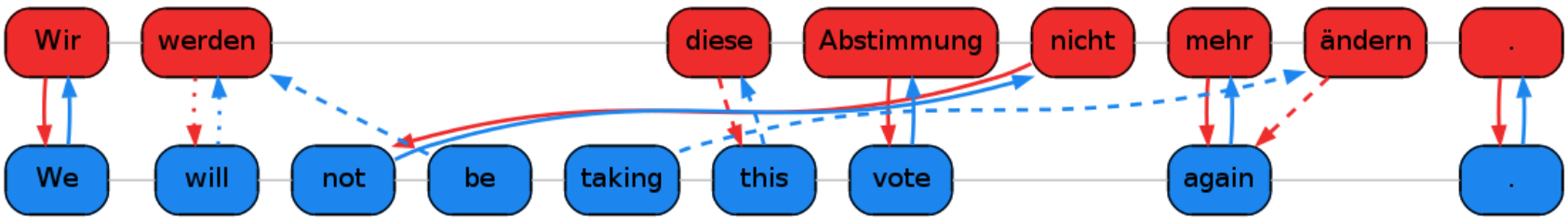
Unser Korpus II

basierend auf 15 Jahren Debatten des Europäischen Parlaments

- mehr als 130.000 Redebeiträge (entspricht 240 Mio. Token)
- in sieben Sprachen: **de**, **en**, **es**, **fr**, **it**, **fi**, **pl**
- Wortarten-Erkennung (*part-of-speech tagging*)
- Lemmatisierung (Reduzierung der flektierten Formen auf eine Grundform, z.B. Infinitiv)
- Satzalignierung (Verknüpfung sich entsprechender Sätze)
- Wortalignierung (Verknüpfung sich entsprechender Token)
- syntaktische Dependenzrelationen

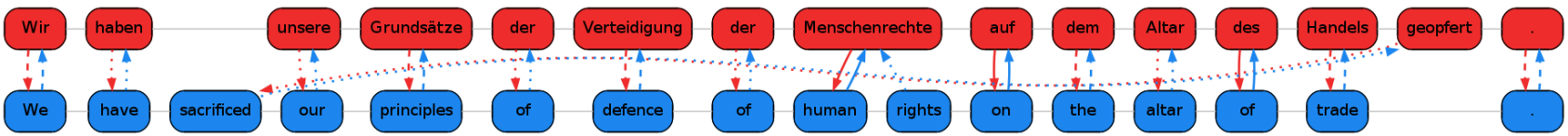
Alignierung

Beispiel 1



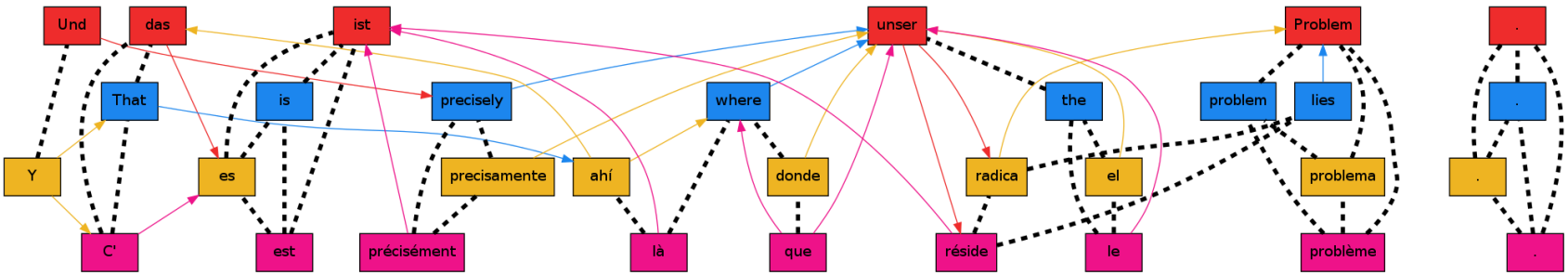
Alignierung

Beispiel 2



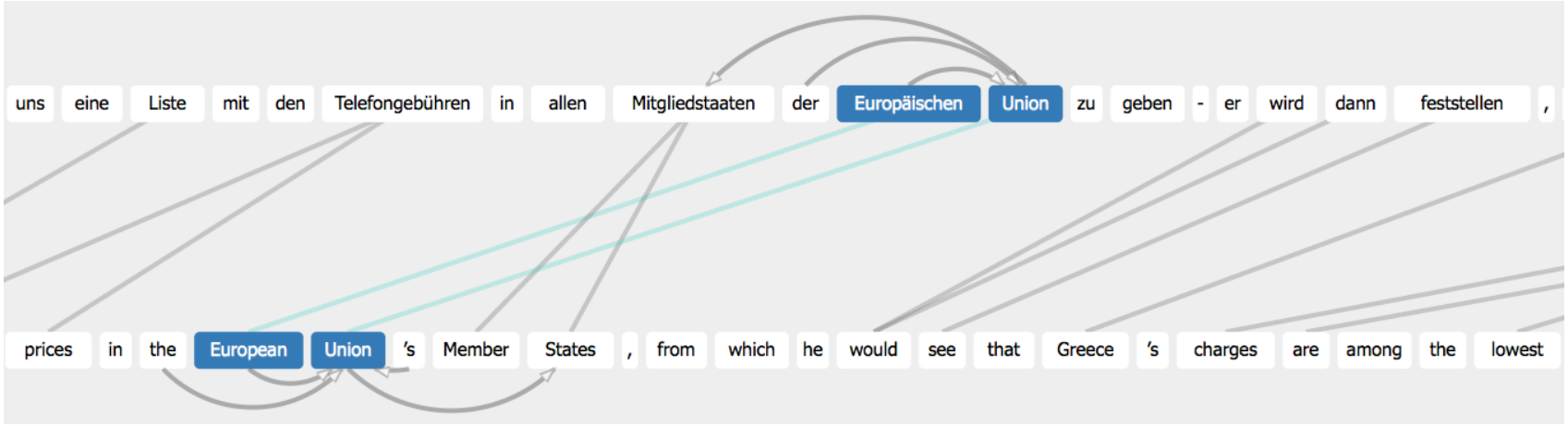
Alignierung

Beispiel 3

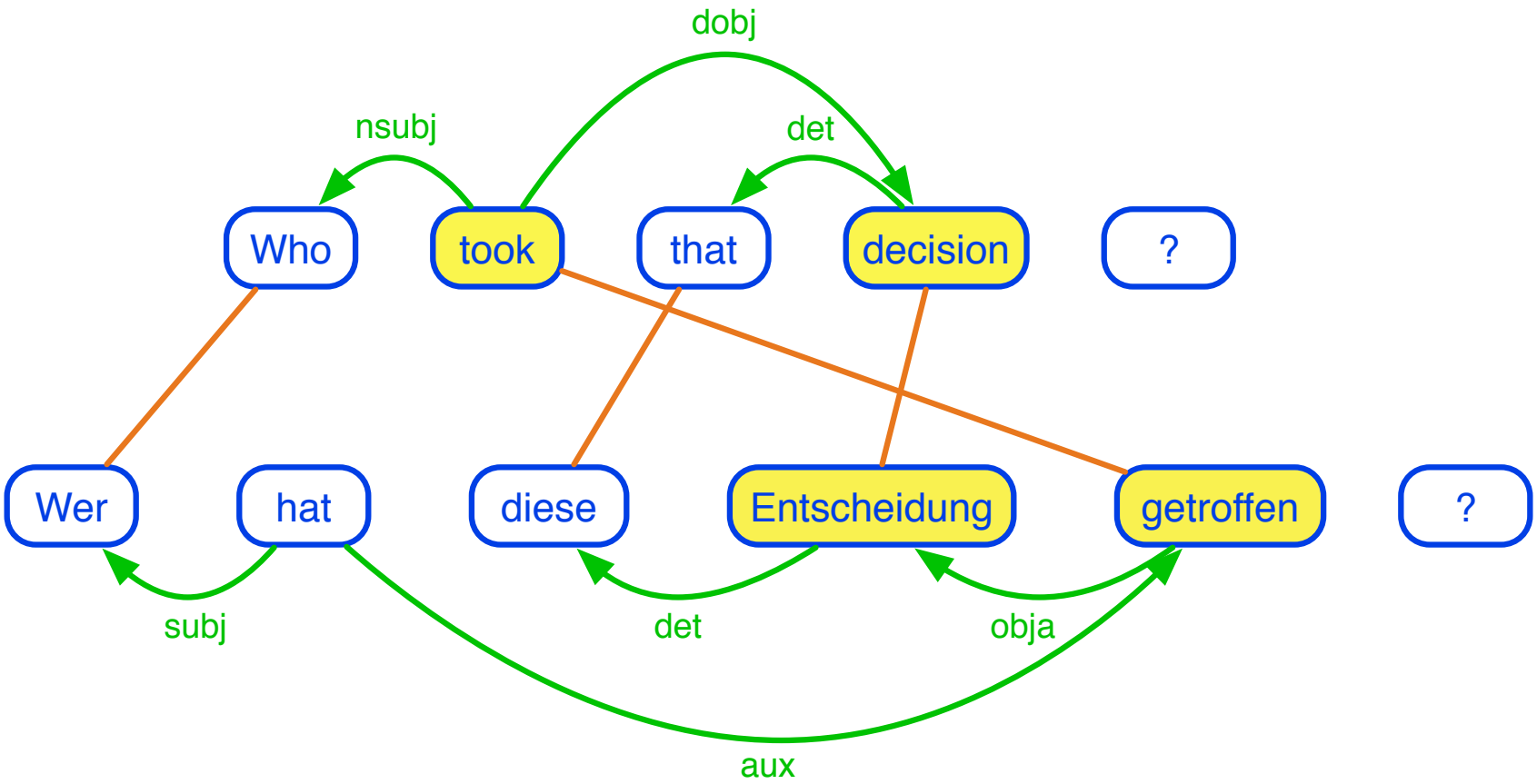


Alignierungen

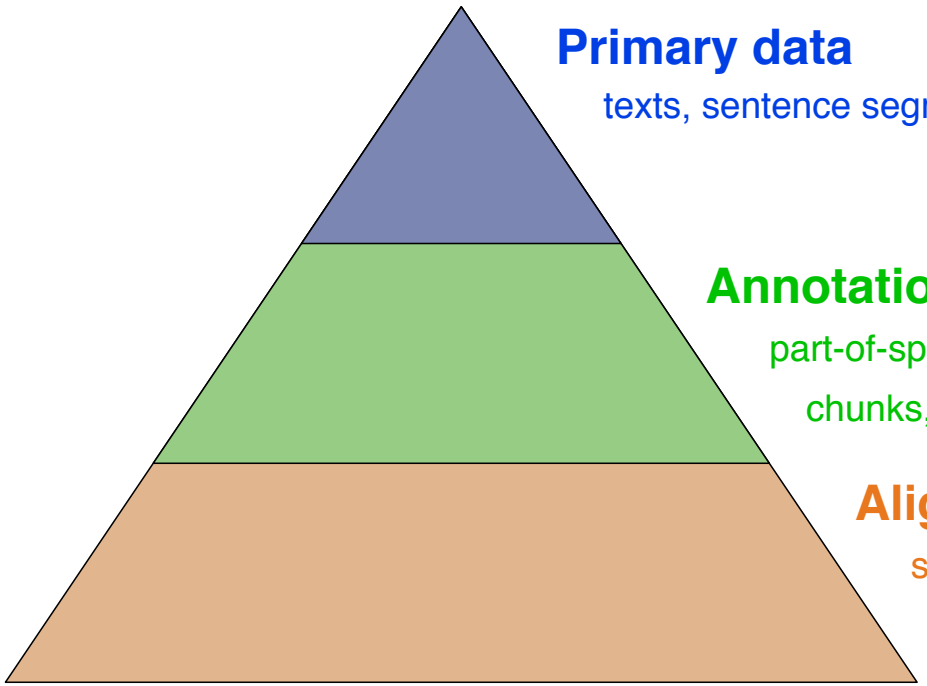
mit syntaktischen Dependenzrelationen



Syntaktische Dependenzrelationen



Korpus-Datenbank



Primary data

texts, sentence segments, tokens

Annotations

part-of-speech tags, lemmas

chunks, dependency parses, coreferences, …

Alignments

sentence and word alignments

multi-parallel hierarchical alignments (MPAs)

Übersicht

Was sind abgetrennte Verbpräfixe?

Example

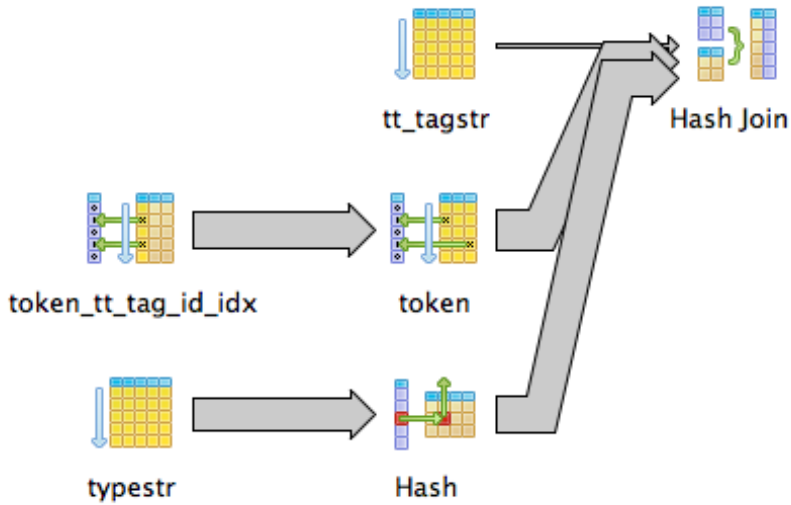
- 70 % der tschechischen Bürger lehnen das System ab.
- Die Schweiz stellt ein viel ernsteres Problem dar.
- Wer die Verhandlung verlässt, stimmt zu.
- Daraufhin schlug er die Unabhängigkeit des Kosovo vor.

Problemstellung

- Zwischen einem Verb und seinem abgetrennten Präfix kann im Deutschen eine grosse Distanz liegen.
- Beim Lemmatisieren wird i.d.R. das Lemma der Verbform (ohne Präfix) erkannt (z.b. 'schlagen' statt 'vorschlagen').
- Wir versuchen algorithmisch das korrekte Lemma zu ermitteln ('vorschlagen'), um die Qualität der Korpusdaten verbessern.
- Zusätzliche Schwierigkeit: In manchen (nicht entscheidbaren) Fällen kommen alternative Lemmata vor:
  'fällt' ... 'aus' $\Rightarrow$ 'ausfallen|ausfällen'

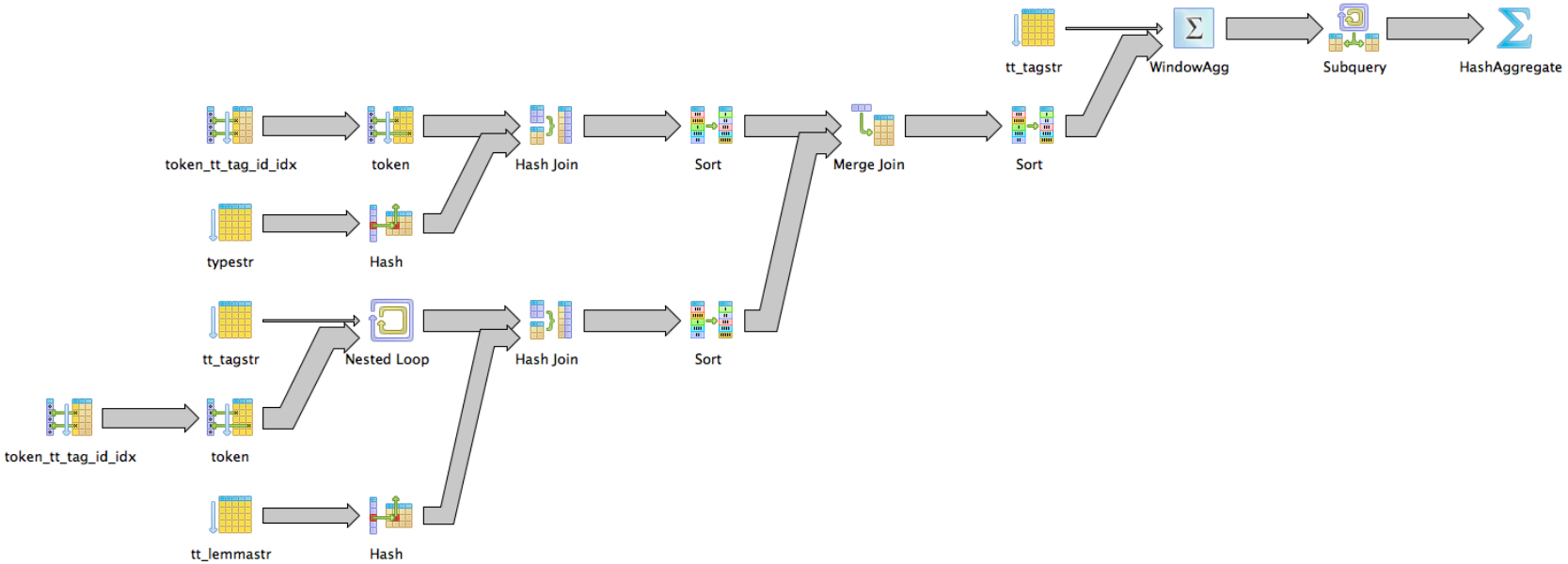
Algorithmus

1. Finde Sätze, in denen abgetrennte Verbpräfixe vorkommen,



Algorithmus II

2. die im linken Kontext ein finites Vollverb aufweisen und
3. fasse diese Verb-Präfix-Paare zusammen.

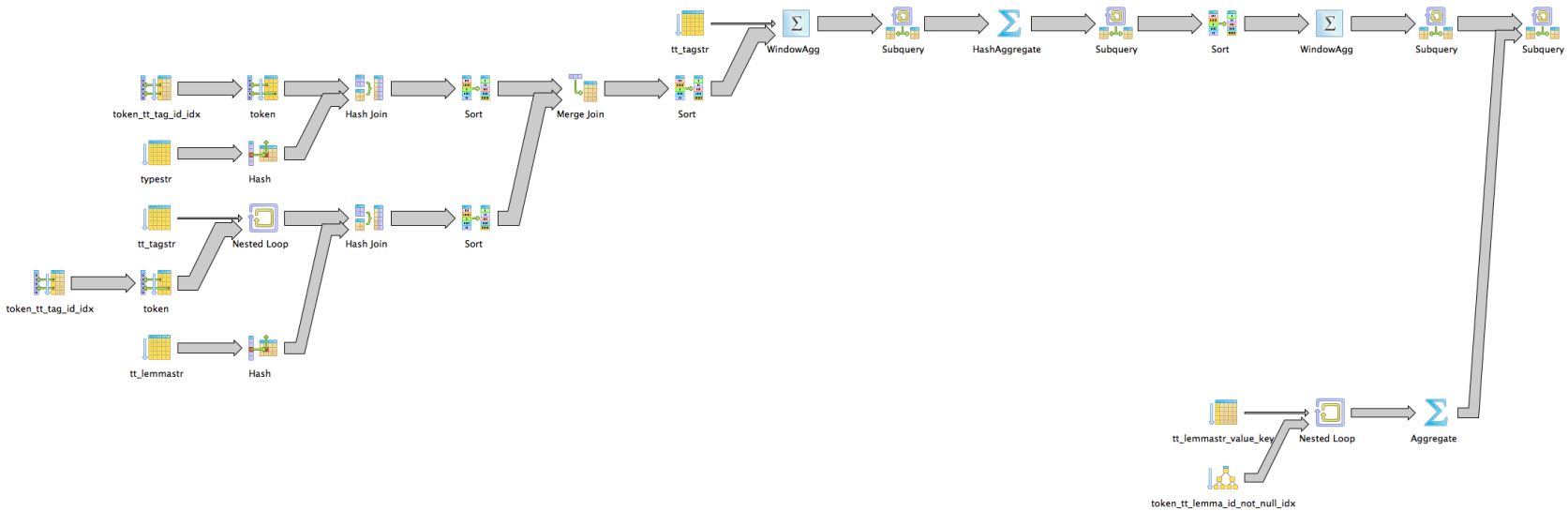


Algorithmus II

```
SELECT vprefix, verb, count(*) AS pv_count
FROM (
    SELECT segment_id, vprefix, tt_lemmastr.val AS verb, row_number()
        OVER (PARTITION BY segment_id ORDER BY token_id DESC) AS pos
    FROM (
        SELECT segment_id, token_id, typestr.val AS vprefix
        FROM token
        JOIN typestr ON typestr.aid = type_id
        WHERE tt_tag_id = (
            SELECT aid
            FROM tt_tagstr
            WHERE val = 'PTKVZ'
            AND language_id = 1550 )
    ) q1
    JOIN token USING (segment_id)
    JOIN tt_tagstr ON tt_tagstr.aid = tt_tag_id
    JOIN tt_lemmastr ON tt_lemmastr.aid = tt_lemma_id
    WHERE tt_tagstr.val IN ('VVFIN','VVIMP')
    AND token.token_id < q1.token_tid
) q2
WHERE pos = 1
GROUP BY vprefix, verb
```

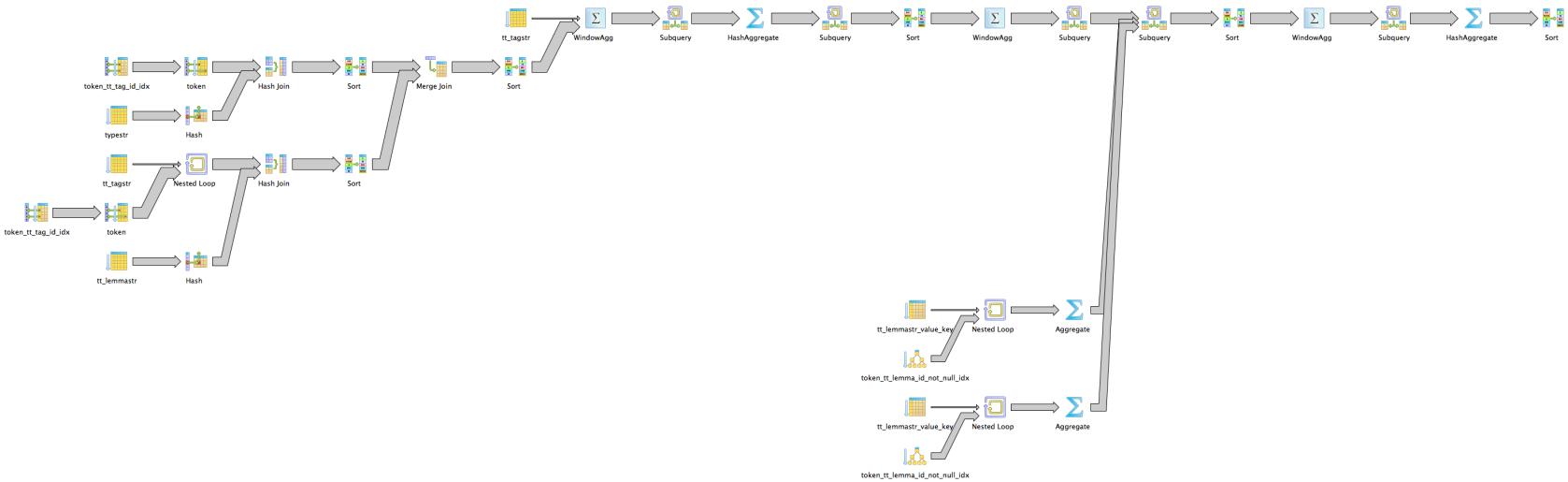
Algorithmus III

4. Bei mehrdeutigen Verblemmata, trenne diese auf (‘führe’ + ‘durch’ $\widehat{=}$ ‘durchfahren’ oder ‘durchführen’).
5. Berechne die Vorkommenshäufigkeiten der komponierten Verblemmata (‘stellt’ + ‘dar’ $\Rightarrow$ ‘darstellen’)



Algorithmus IV

6. Pro Verb-Präfix-Paar wähle dasjenige komponierte Verblemma, das am häufigsten vorkommt (min. 1 Beleg).



Algorithmus IV

```
SELECT vprefix, verb, sum(pv_count) AS pv_count
FROM (
    SELECT *, row_number() OVER (PARTITION BY id ORDER BY l_count DESC) l_pos
    FROM (
        SELECT *, (
                SELECT count(*) AS l_count
                FROM token
                JOIN tt_lemmastr ON aid = tt_lemma_id
                WHERE val = vprefix||verb )
        FROM (
            SELECT id, pv_count, vprefix,
                unnest(regexp_split_to_array(verb,E'\|')) AS verb
            FROM (
                SELECT *, row_number() OVER (ORDER BY pv_count DESC) AS id
                    FROM (...) q3
            ) q4
        ) q5
    ) q6
    WHERE l_count > 0
) q7
WHERE l_pos = 1
GROUP BY vprefix, verb
ORDER BY pv_count DESC
```

Ergebnis

Präfix	Verb	Anzahl
dar	stellen	6903
statt	finden	4868
auf	fordern	4789
zu	stimmen	4372
vor	schlagen	4201
fest	stellen	2537
vor	liegen	1888
vor	sehen	1874
überein	stimmen	1861
ab	lehnen	1688
aus	gehen	1618
hin	weisen	1608
bei	tragen	1520
an	schließen	1419
⋮		

Übersicht

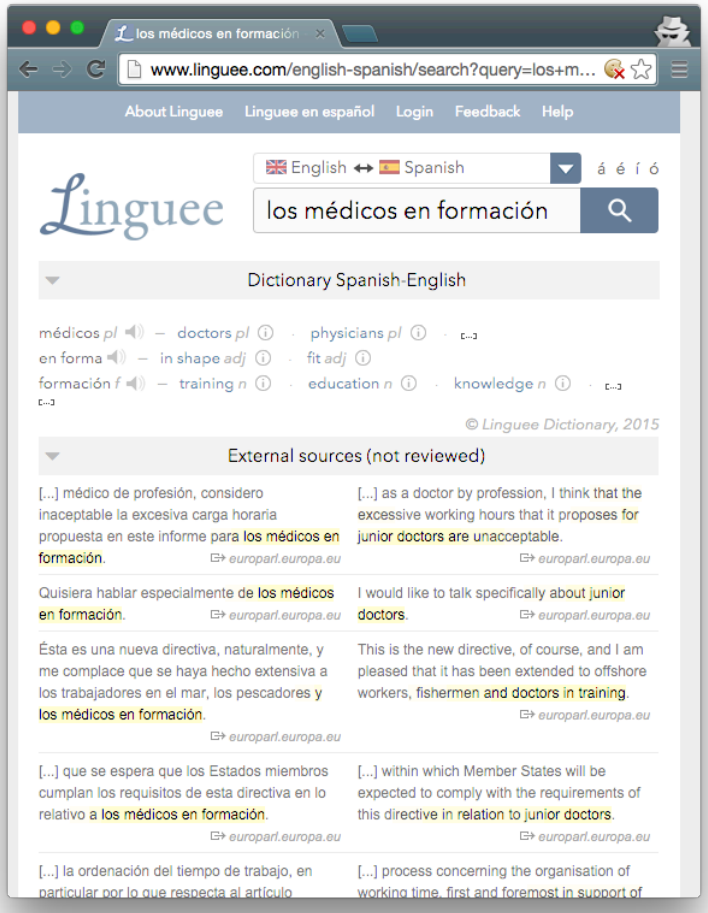
Motivation

- Mehrwortausdrücke stellen für Übersetzer und Sprachenlerner eine besondere Herausforderung dar.
- Grosse parallele Korpora enthalten viele Beispielübersetzungen.
- Es gibt bereits Webanwendungen, mit deren Hilfe der Benutzer Übersetzungen in parallelen Korpora suchen kann (Glosbe[1], Linguee[2], Tradooit[3], ...),
- die Suche ist jedoch jeweils auf ein Sprachpaar beschränkt.

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

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
ä ö ü ß
English ↔ German
English → German
German → English
English ↔ Portuguese
English ↔ Spanish
English ↔ French
English ↔ Italian
English ↔ Russian
English ↔ Japanese
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

Multilingwis

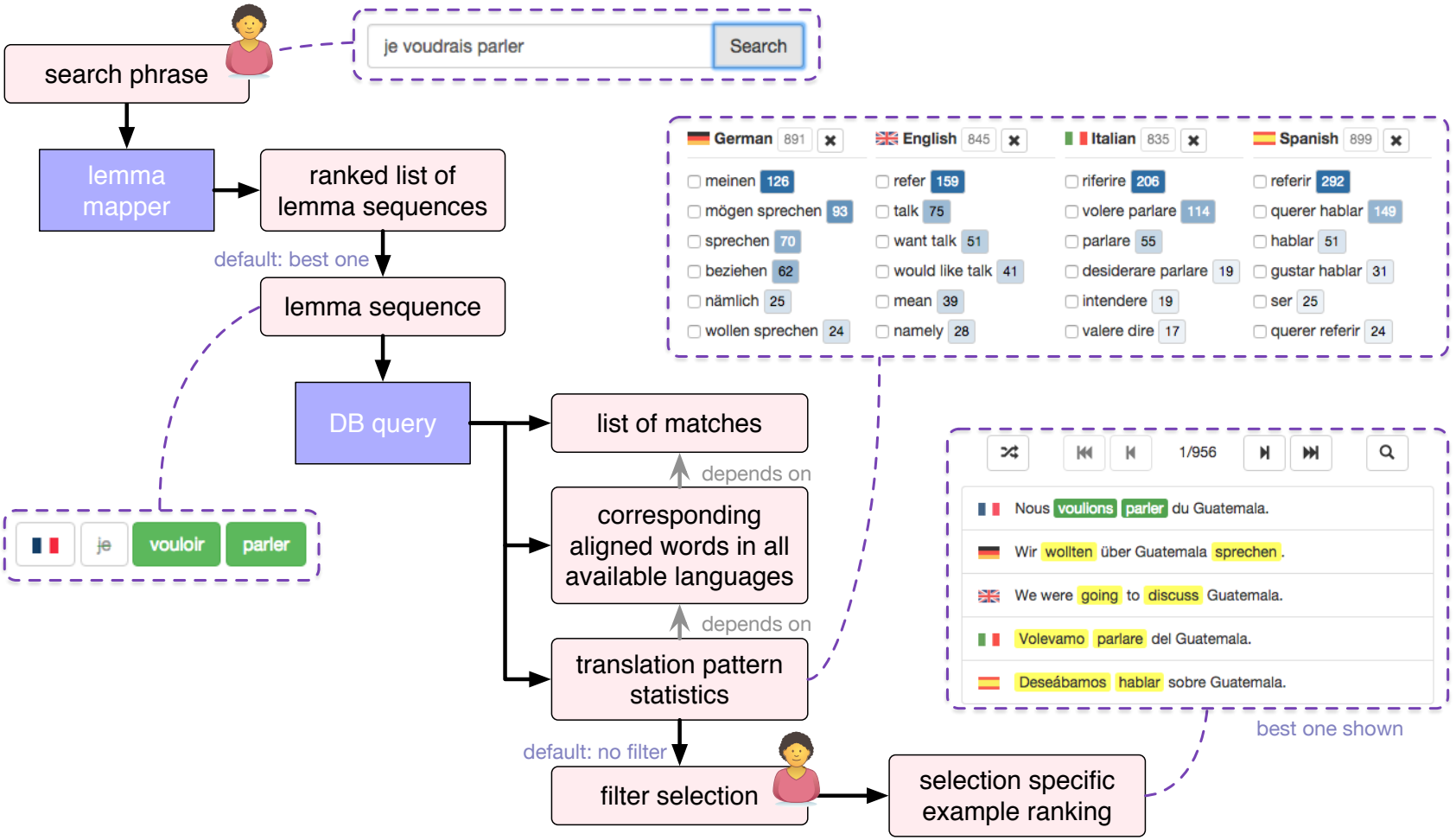
Multilingwis

Multilingual Word Information System

Multilingwis ...

- ist ein Tool zur Exploration von Übersetzungsvarianten,
- sucht mehrsprachig (z.Zt. **de**, **en**, **es**, **fr**, **it**),
- bietet eine Rückwärtssuche zu jeder Übersetzungsvariante,
- zeigt Beispiele mit markierten Übersetzungsäquivalenten.

Ablauf

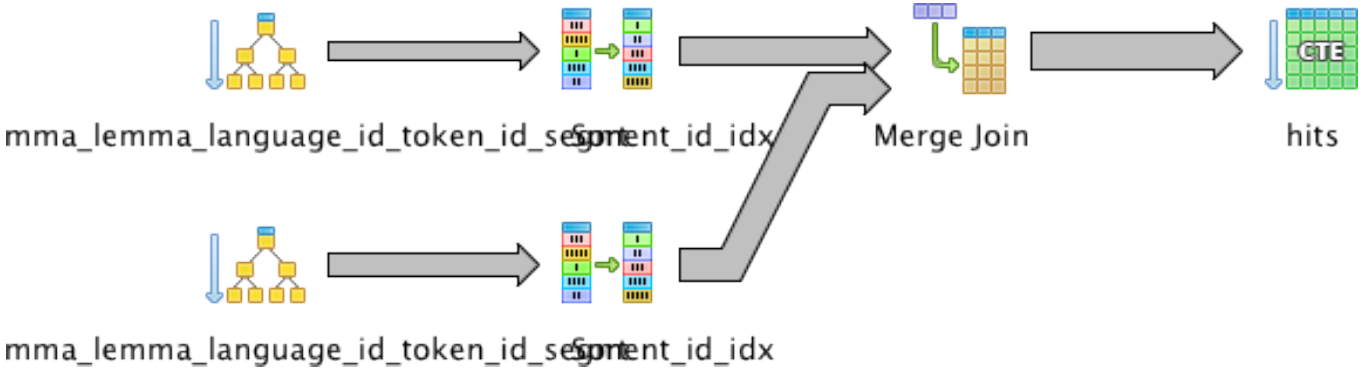

search phrase
je voudrais parler
Search
lemma mapper
ranked list of lemma sequences
default: best one
lemma sequence
DB query
list of matches
depends on
corresponding aligned words in all available languages
depends on
translation pattern statistics
default: no filter
filter selection
selection specific example ranking
best one shown
je
vouloir
parler
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
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.

Demo

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

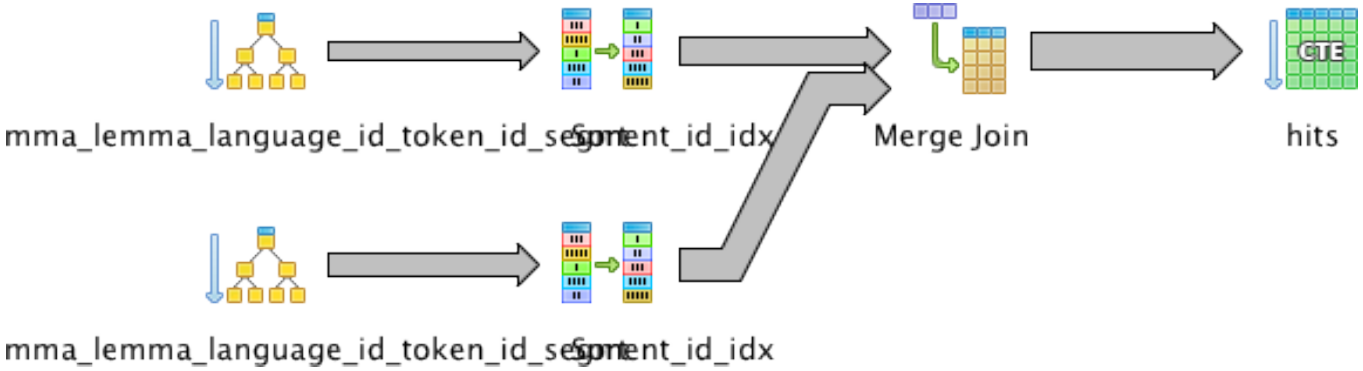
Suchtreffer – Query Plan

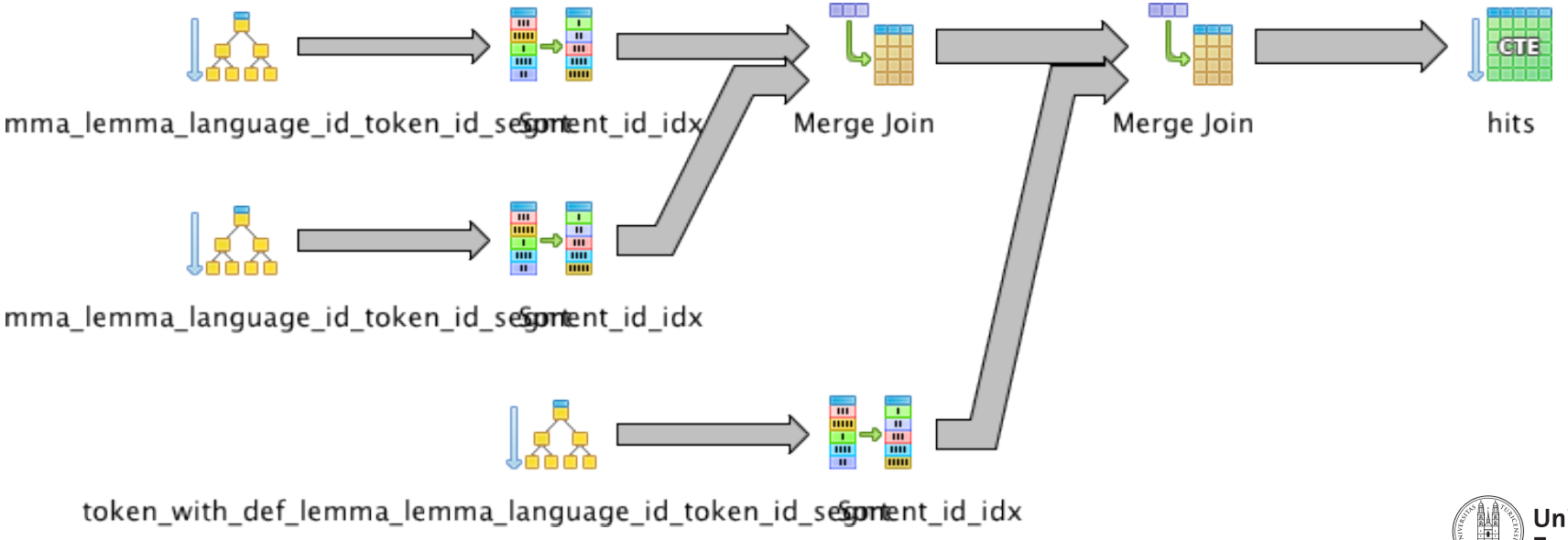
für Lemmapaare und -tripel



Suchtreffer – Query Plan

für Lemmapaare und -tripel



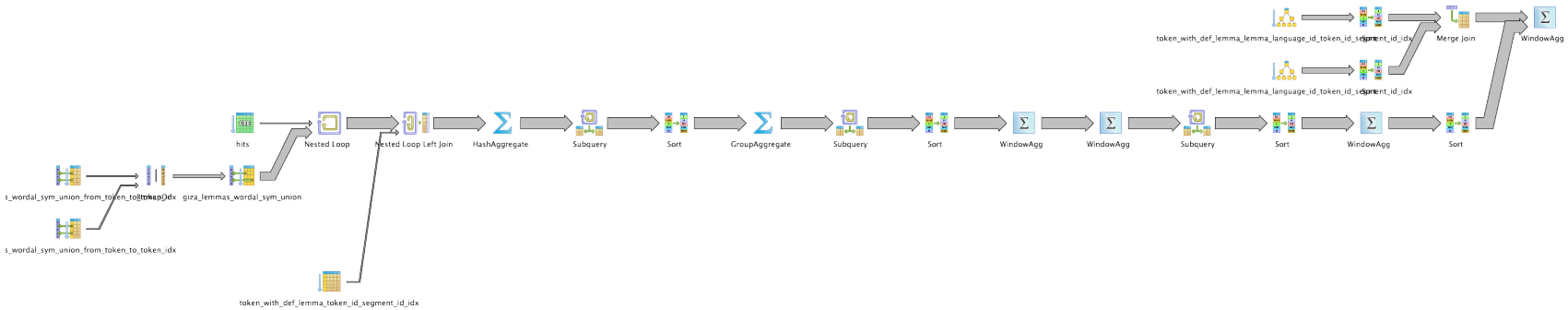


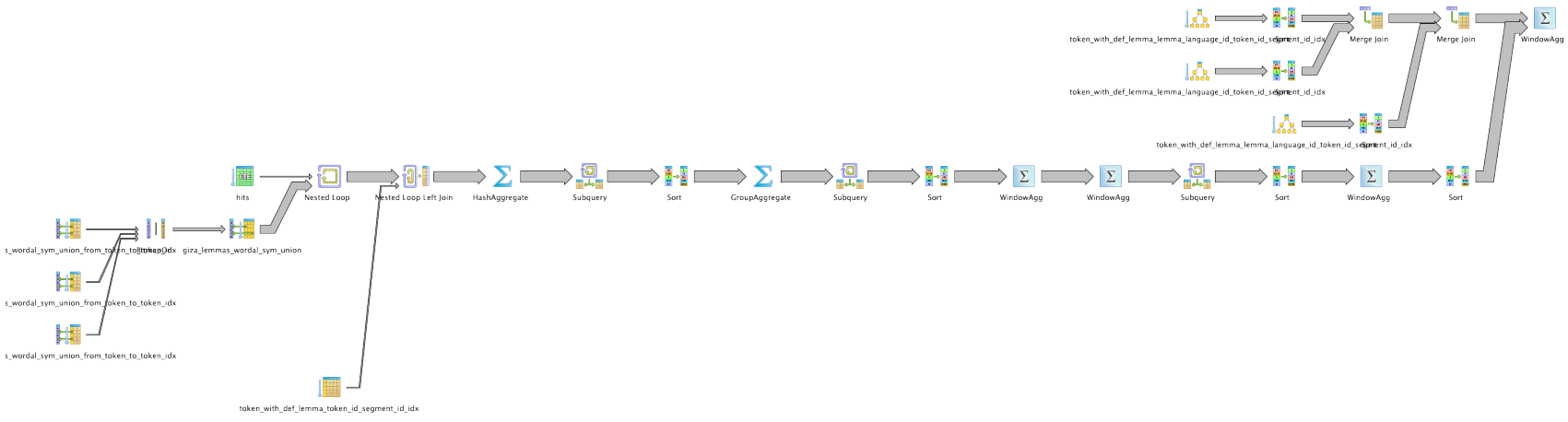
Suchtreffer – Query

```
WITH hits AS (
    SELECT *
    FROM (
        SELECT s1.token_id t1, s2.token_id t2, s1.segment_id AS s_sid, (
            SELECT COALESCE(bool_or(upos IN ('NOUN','VERB','ADJ','ADV')),false)
            FROM fullep.token
            LEFT JOIN fullep.tt_tagstr ON aid = tt_tag_id
            WHERE token_id BETWEEN s1.token_id+1 AND s2.token_id-1 ) has_cw1
        FROM (
            SELECT *
            FROM token_with_def_lemma
            WHERE lemma = $2
        ) s1
        JOIN (
            SELECT *
            FROM token_with_def_lemma
            WHERE lemma = $3
        ) s2
        ON s2.segment_id = s1.segment_id
            AND s2.token_id BETWEEN s1.token_id+1 AND s1.token_id+4
        WHERE $1 IS NULL OR s1.language_id = $1
    ) x
    WHERE NOT has_cw1
)
```

Übersetzungsäquivalente – Query Plan

für Lemmapaare und -tripel



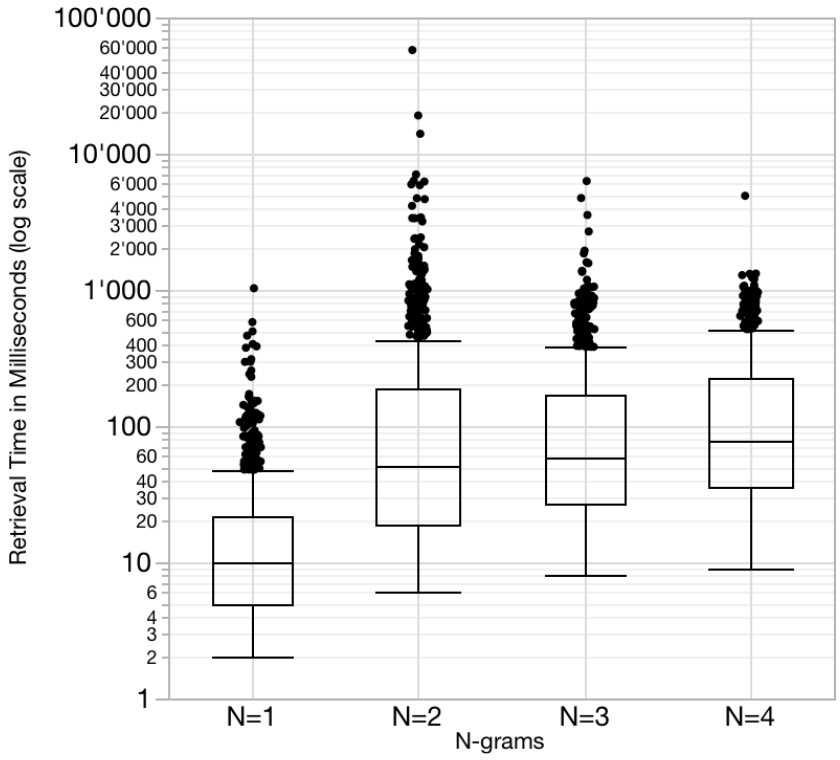


Übersetzungsäquivalente – Query

```
SELECT ...
FROM (
    SELECT x3.*, (SELECT count(*) FROM fullep.token WHERE sid = t_sid) score,
        count(*) OVER (PARTITION BY language_id, t_lemmas) lemma_count,
        dense_rank() OVER (ORDER BY language_id, t_lemmas) cluster_id
    FROM (
        SELECT s_sid, s_tids, language_id, t_sid,
            array_agg(t_tid ORDER BY t_tid) t_tids,
            array_agg(t_lemma ORDER BY t_tid) t_lemmas
        FROM (
            SELECT DISTINCT x1.s_sid, ARRAY[t1,t2] s_tids, x1.t_tid,
                twl.sid t_sid, twl.lemma t_lemma, twl.language_id
            FROM (
                SELECT hits.*, to_token t_tid
                FROM hits
                JOIN fullep.berk_lemmas_wordal_sym ON from_token IN (t1,t2)
            ) x1
            LEFT JOIN mlw.token_wdef_lemma twl ON twl.tid = x1.t_tid
        ) x2
        GROUP BY s_sid, s_tids, language_id, t_sid
        HAVING count(DISTINCT t_sid) = 1
    ) x3
) x4
```

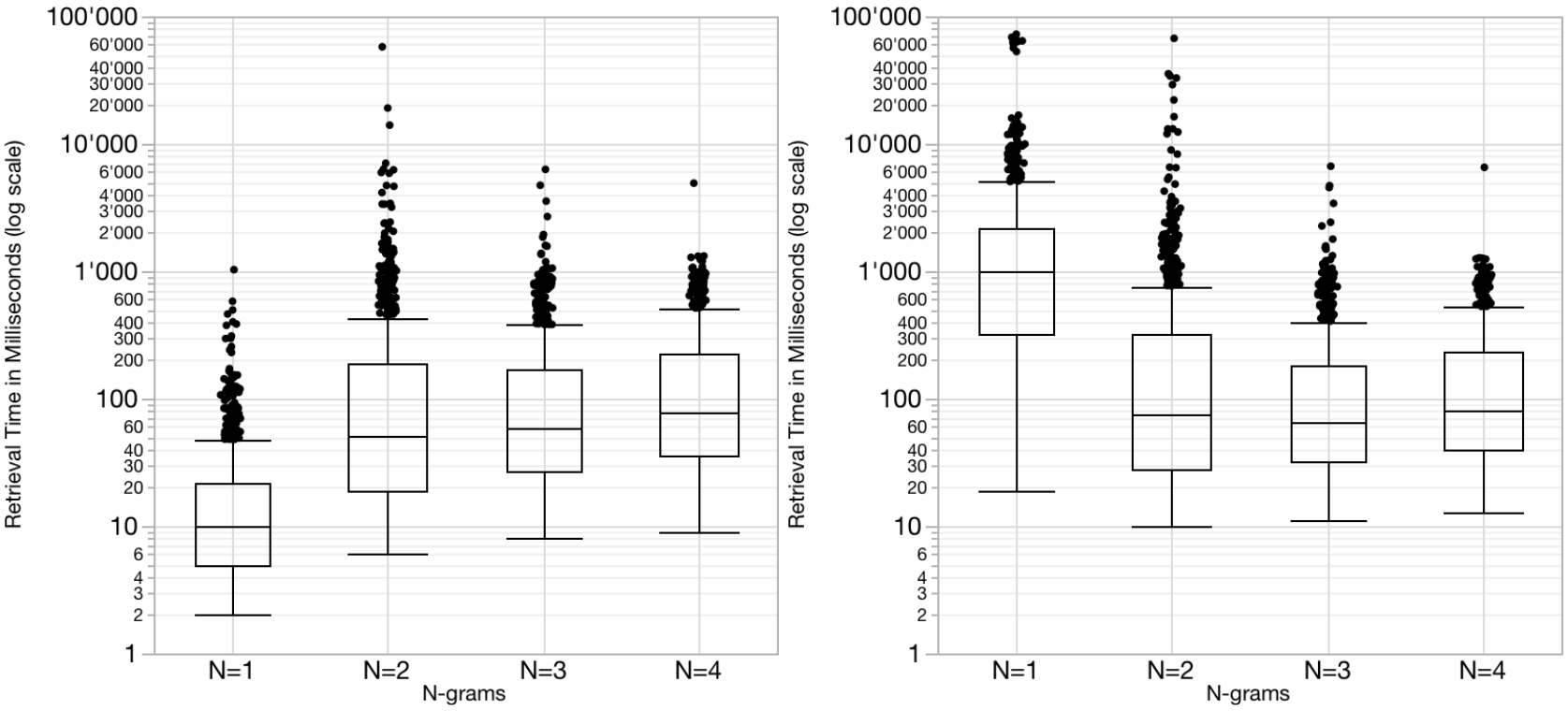
Performance

Suchtreffer (links) and Übersetzungsäquivalente (rechts)



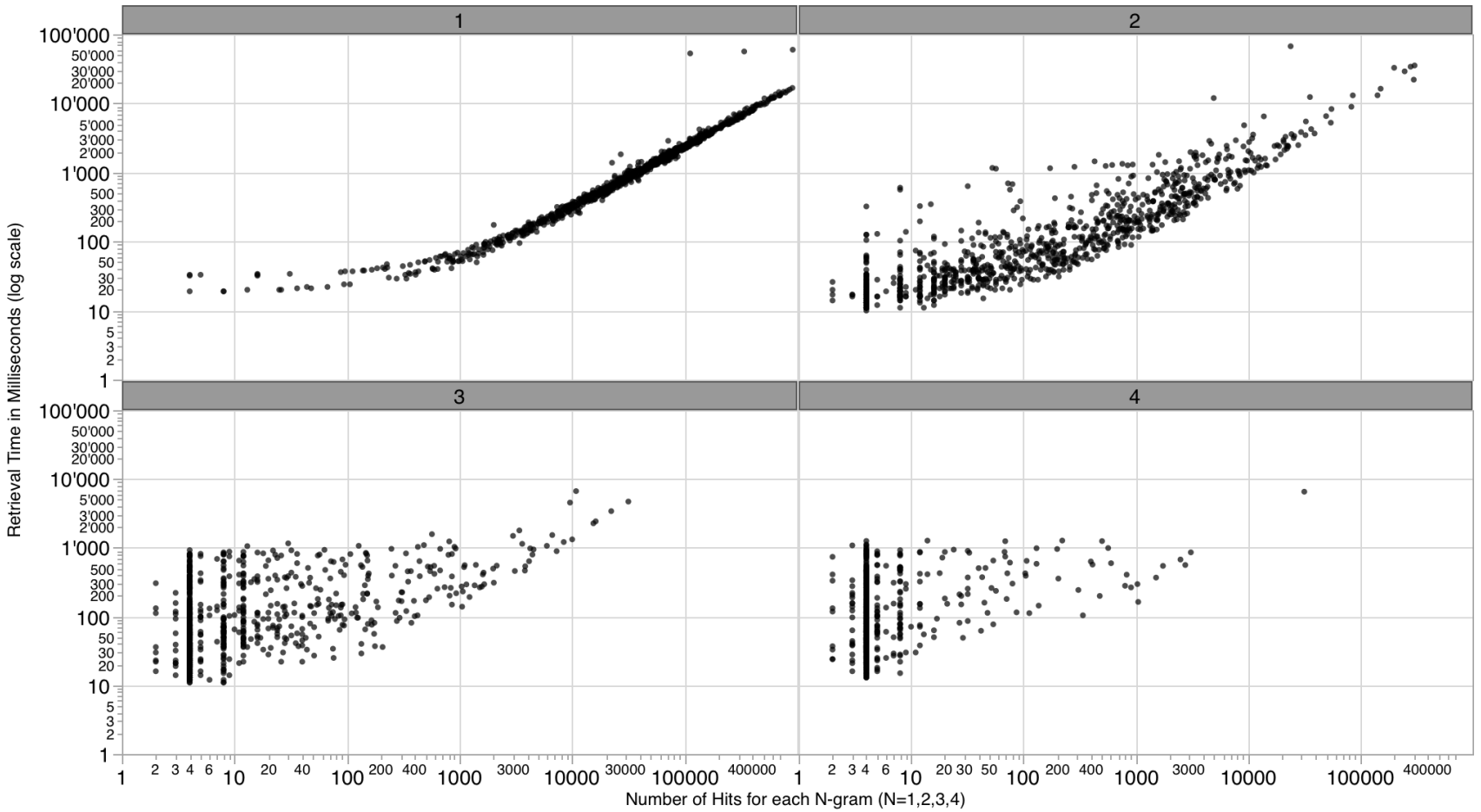
Performance

Suchtreffer (links) and Übersetzungsäquivalente (rechts)



Performance

Korrelation der Anzahl Übersetzungsäquivalente zur Suchzeit



Übersicht

Geokollokationen

N-Gramme

nGrams-Viewer

Korpusauswahl bundestagwp17v2linke_ngrams.txt.new.sorted.tab (1753 nGrams)

idNgram	ngram	pValue	nrOfInsta
18	ADJD verstanden VVPP , \$, dass KOUS Sie PPER	0.05	12
19	Sie PPER ADJD verstanden VVPP , \$, dass KOUS Sie PPER	0.01	9
20	Sie PPER ADJD verstanden VVPP , \$, dass KOUS	0.05	12
21	verstanden VVPP , \$, dass KOUS Sie PPER ART	0.05	7
22	ADJD verstanden VVPP , \$, dass KOUS Sie PPER ART	0.05	7
23	Sie PPER ADJD verstanden VVPP , \$, dass KOUS Sie PPER ART	0.05	5
24	Wenn KOUS ich PPER Sie PPER ADJD verstanden VVPP habe VAFIN , \$,	0.05	6
25	Wenn KOUS ich PPER Sie PPER ADJD verstanden VVPP habe VAFIN	0.05	6
26	ich PPER Sie PPER ADJD verstanden VVPP	0.01	26
27	Sie PPER ADJD verstanden VVPP	0.05	28
28	Habe VAFIN ich PPER Sie PPER ADJD verstanden VVPP , \$, dass KOUS	0.01	9
29	Habe VAFIN ich PPER Sie PPER ADJD verstanden VVPP , \$,	0.05	9
30	Habe VAFIN ich PPER Sie PPER ADJD verstanden VVPP	0.05	9
31	Erklären VVFIN Sie PPER das PDS 11 x Erklären Sie das	0.05	11
32	Erklären VVFIN Sie PPER das PDS ADV	0.05	7
33	Erklären VVFIN Sie PPER das PDS ADV ART	0.05	5
34	Erklären VVFIN Sie PPER ADV ADV	0.05	10
35	Erklären VVFIN Sie PPER mir PRF ADV ADV 5 x Erklären Sie mir doch einmal 1 x Erklären Sie mir doch jetzt 1 x Erklären Sie mir doch bitte 1 x Erklären Sie mir dann aber	0.01	8
36	Erklären VVFIN Sie PPER mir PRF ADV	0.05	
37	ADV sagen VVFIN Sie PPER , \$,	0.05	

2011 2011-01-27 DIE_LINKE 87 Dr. Gregor Gysi Abgeordnete/r 2011-01
weil beide am Arbeitsmarkt flexibel sein sollen . **Erklären Sie mir doch einmal** , wie Ihre dritte These dazu passt :

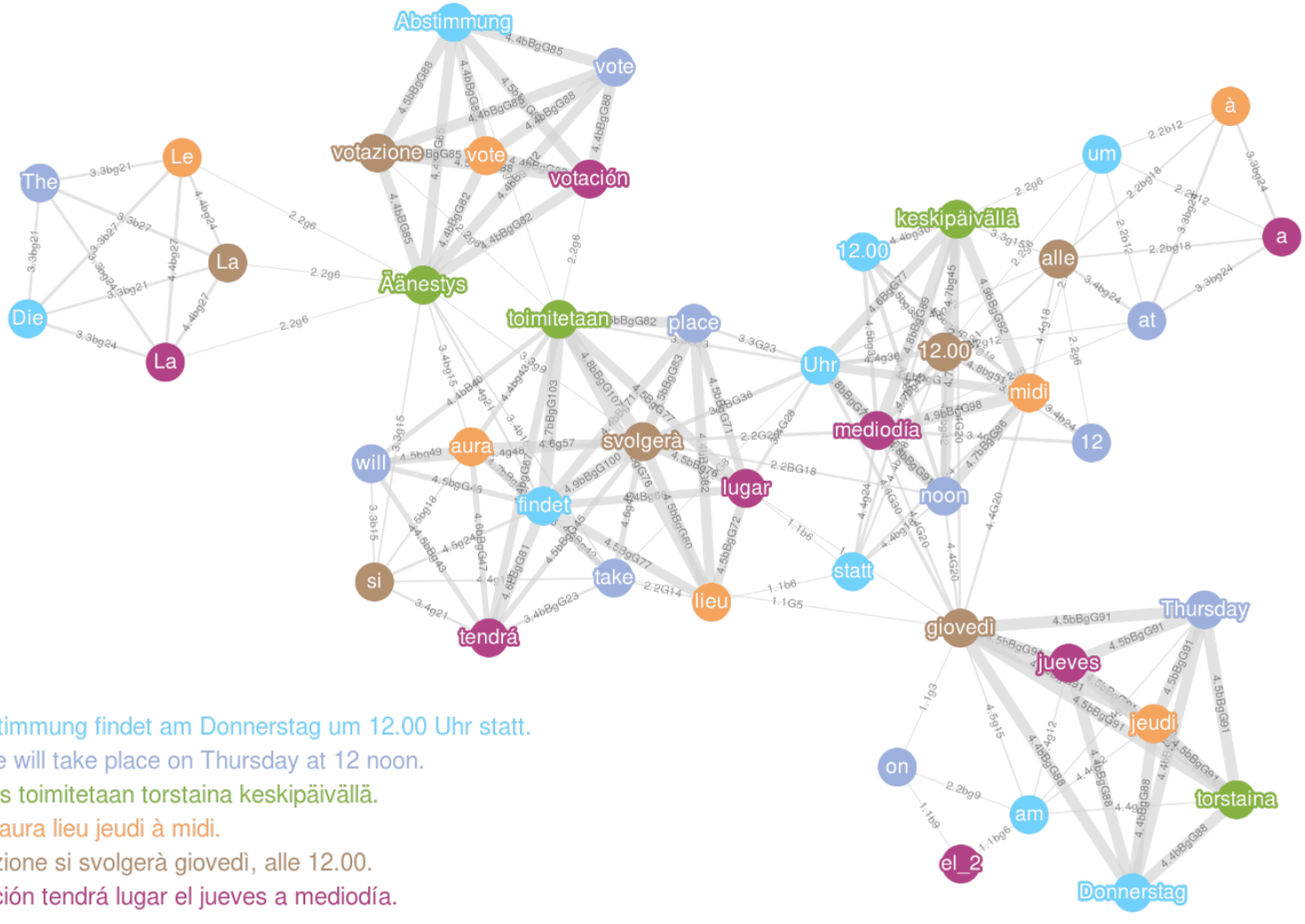
2012 2012-03-29 DIE_LINKE 172 Dr. Gregor Gysi Abgeordnete/r 2012-03
-Kauder , aber nicht mit Ihrer Politik . **Erklären Sie mir doch einmal** , warum die Europäische Zentralbank übrigens auch

2011 2011-07-06 DIE_LINKE 119 Dr. Gregor Gysi Abgeordnete/r 2011-07
daraus machen . Was haben wir davon ? **Erklären Sie mir doch einmal** : Was haben wir davon , außer dass

2011 2011-09-08 DIE_LINKE 124 Klaus Ernst Abgeordnete/r 2011-09
ich zitiere - ' außenwirtschaftliches Gleichgewicht ' . **Erklären Sie mir doch einmal** - Frau Merkel ist ja nicht mehr da

2009 2009-11-10 DIE_LINKE 3 Wolfgang Gehrcke 2009-11
' die Idee der westlichen Werte ist . **Erklären Sie mir doch einmal** , was für Sie die westlichen Werte sind

Alignierungsgraph



Die Abstimmung findet am Donnerstag um 12.00 Uhr statt.
The vote will take place on Thursday at 12 noon.
Äänestys toimitetaan torstaina keskipäivällä.
Le vote aura lieu jeudi à midi.
La votazione si svolgerà giovedì, alle 12.00.
La votación tendrá lugar el jueves a mediodía.

Übersicht

Ausblick

Technisch:

- DBen auf zweiten Server replizieren; Lastverteilung mit pgpool-II/pgBouncer
- Schnellere Anfragen durch parallele Berechnung (9.6?)

Computerlinguistisch:

- Multilingwis in 7 Sprachen; erweiterte Suche (Wortarten, syntaktische Struktur, ...)
- Automatische Konvertierung linguistischer Anfragen in (performanten) SQL-Code

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: