University of Zurich[UZH]

Institute of Computational Linguistics

Verfahren zur sprachübergreifenden Phrasemsuche in dependenzannotierten, alignierten Korpora

Johannes Graën — LeKo, Innsbruck, 12.02.2016

Übersicht

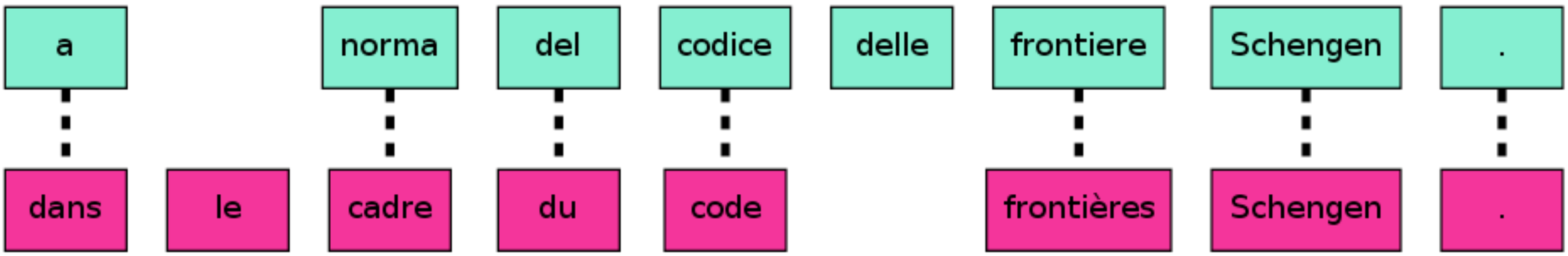
- Vorstellung (Projekt, Schwerpunkt)
- Korpus (Quelle, Aufbau, Struktur)
- Korpus-Anfragen (Beispiele)
- Ausblick

Vorstellung

Vorstellung

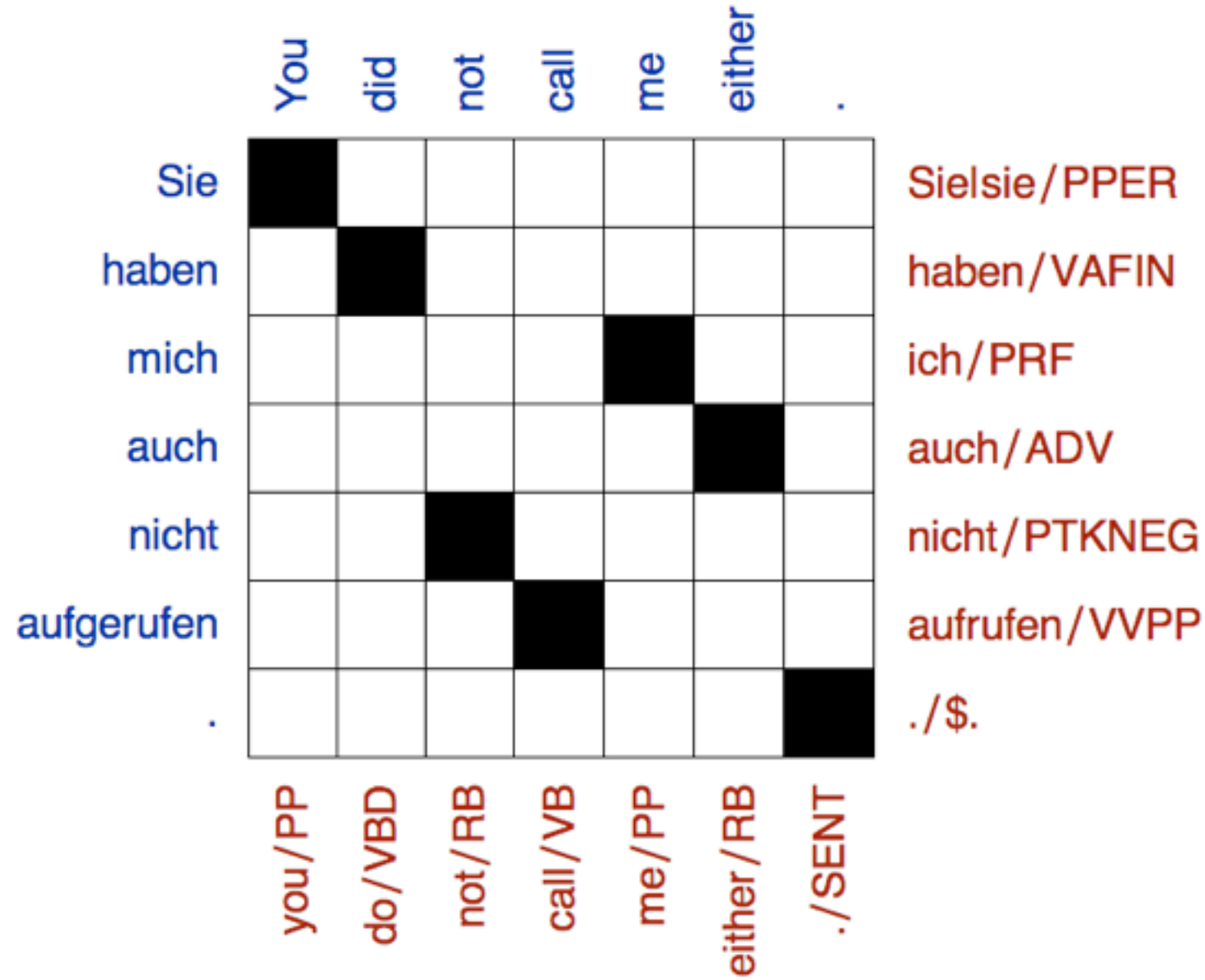
- **SPARCLING**
  (large-scale parallel corpora to study linguistic variation)

  - variabler Artikelgebrauch:



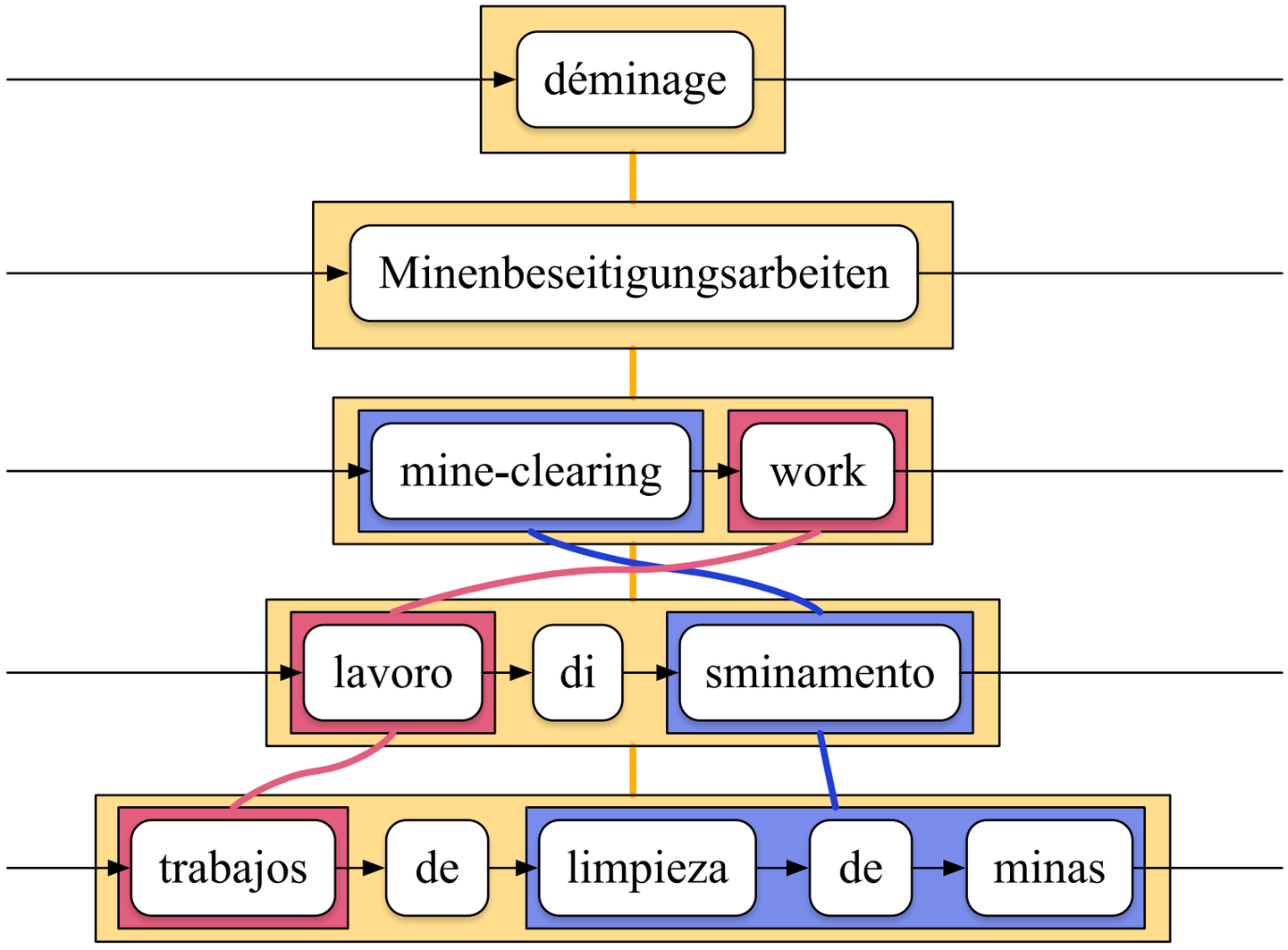
Vorstellung

- Wort-Alignierungen:



Vorstellung

- Multi-parallele Alignierungen:



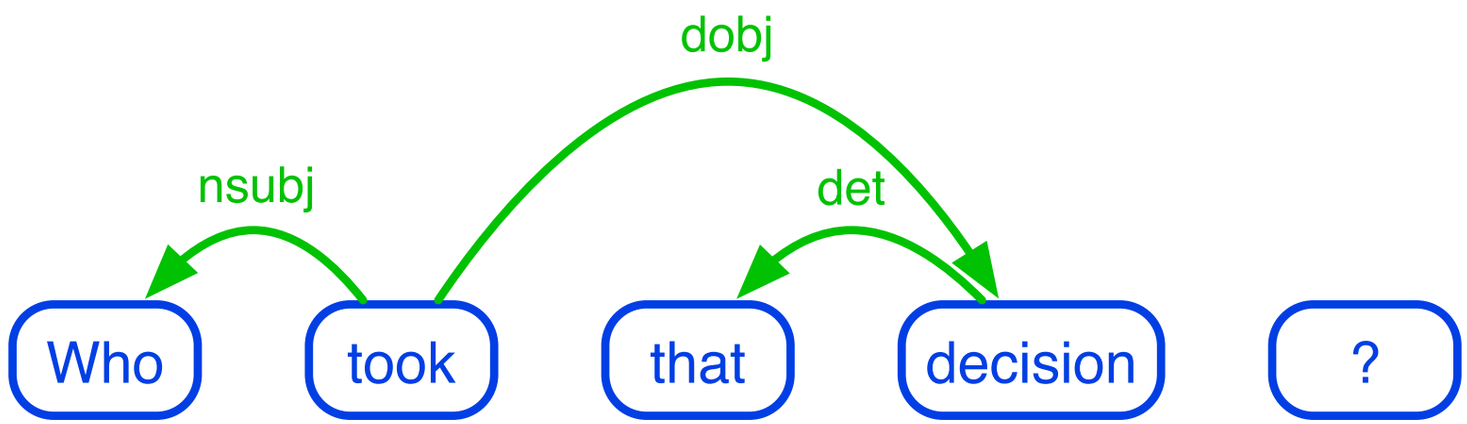
Korpus

Korpus

- Anforderungen
    - große Datenmenge
        - verschiedene Sprachen/Sprachfamilien
            - mindestens Deutsch und Englisch
- *Corrected & Structured Europarl Corpus (CoStEP)*
    - http://pub.cl.uzh.ch/purl/costep

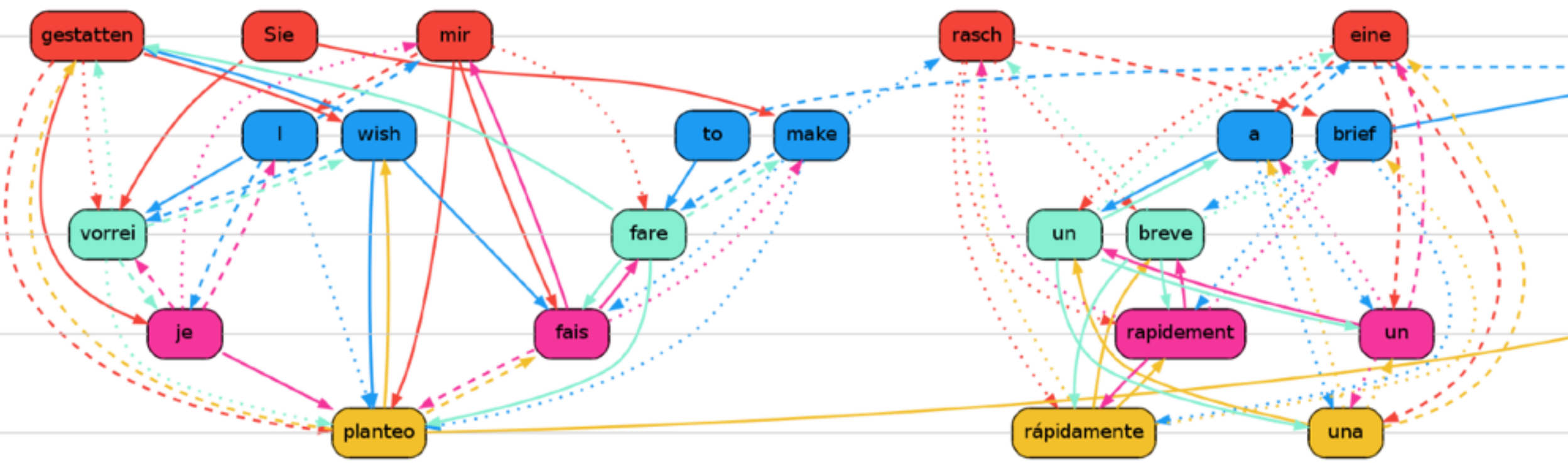
Korpus

- Annotation
  - PoS-Tagging (und universelle PoS-Tags), Lemmatisierung, Dependenz-Parsing

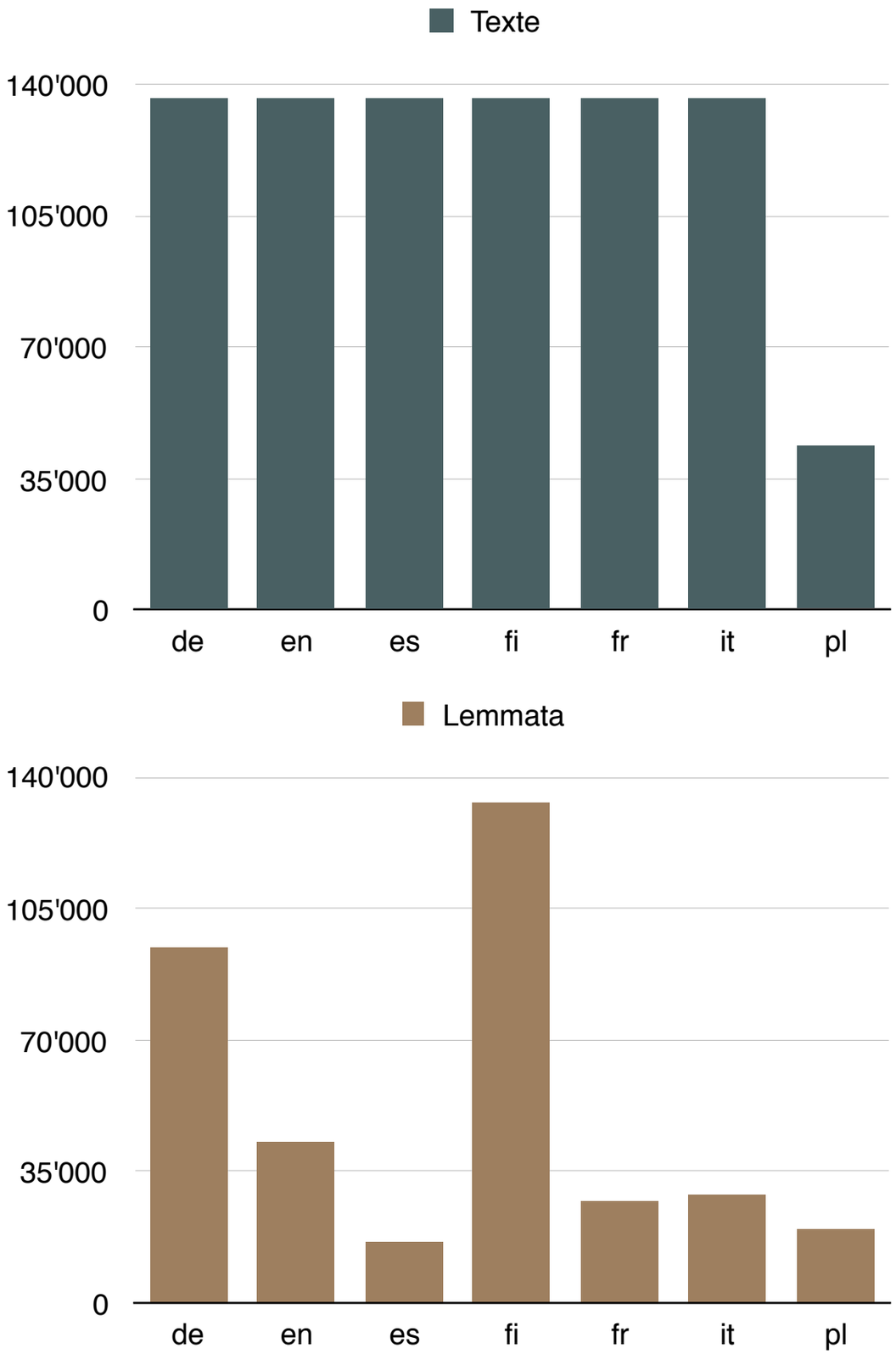


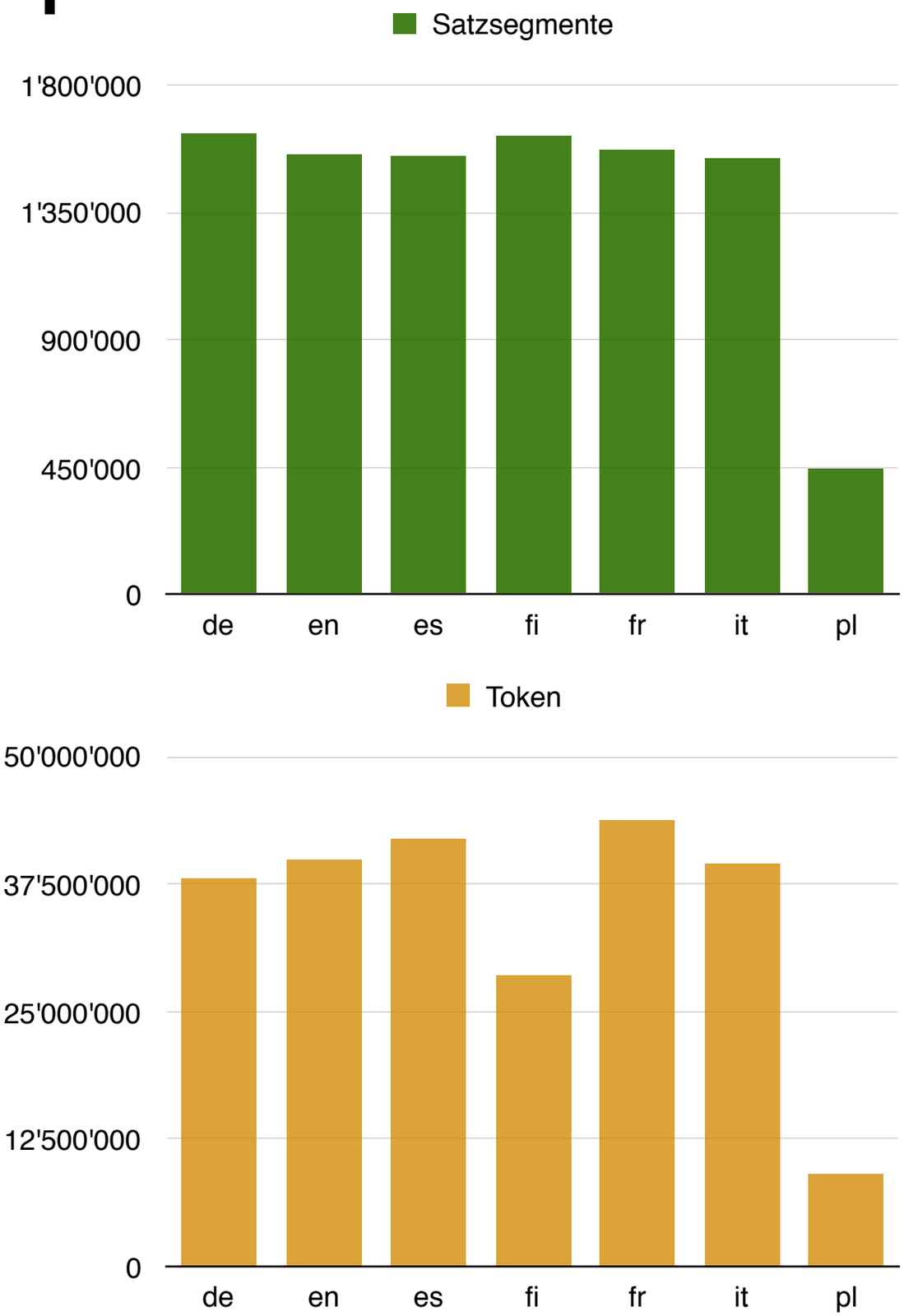
Korpus

- Alignierung
  - Texte (in CoStEP), Satzsegmente, Wörter/Token



Korpus


Texte
140'000
105'000
70'000
35'000
0
de
en
es
fi
fr
it
pl


Satzsegmente
1'800'000
1'350'000
900'000
450'000
0
de
en
es
fi
fr
it
pl

Lemmata
140'000
105'000
70'000
35'000
0
de
en
es
fi
fr
it
pl

Token
50'000'000
37'500'000
25'000'000
12'500'000
0
de
en
es
fi
fr
it
pl

Korpus-Anfragen

(1) Lemma-Korrektur für Verben mit abgetrennten Verbpräfixen

	1	2	3	4	5	6	7	8	9	10	11	12
F	Dies	stellt	im	eigentlichen	Sinn	die	größte	Herausforderung	von	allen	dar	.
L	dies	stellen	in	eigentlich	Sinn	die	groß	Herausforderung	von	alle	dar	.
T	PDS	VVFIN	APPRART	ADJA	NN	ART	ADJA	NN	APPR	PIS	PTKVZ	$.

	1	2	3	4	5	6	7	8	9	10
F	70	%	der	tschechischen	Bürger	lehnen	das	System	ab	.
L	70	%	die	tschechisch	Bürger	lehnen	die	System	ab	.
T	CARD	NN	ART	ADJA	NN	VVFIN	ART	NN	PTKVZ	$.

Präfix	Verblemma	Anzahl
dar	stellen	6910
statt	finden	4872
auf	fordern	4792
zu	stimmen	4383
vor	schlagen	4200
fest	stellen	2538
vor	liegen	1891
vor	sehen	1873
überein	stimmen	1864
ab	lehnen	1690
aus	gehen	1617
hin	weisen	1606
bei	tragen	1523
an	schließen	1425
aus	reichen	1361
aus	sehen	1291
ab	zielen	1250
ab	hängen	1202
an	nehmen	1086
an	kommen	1057
auf	rufen	1049
aus	sprechen	1010

Vorgehen

1. Finde zu einem Verbpräfix (PTKVZ) das nächstgelegene finite Verb (VVFIN oder VVIMP) im linken Satzkontext.

2. Überprüfe die Kombination Präfix + Verblemma auf Existenz (bei ambigen Lemmata entscheide nach Häufigkeit).

Präfix	Verblemma	Anzahl	en	it
dar	stellen	6910	be (59.2%); represent (15.0%)	rappresentare (30.0%); essere (26.3%); costituire (23.0%)
statt	finden	4872	will,take (58.6%); will (15.7%)	svolgere (81.4%)
auf	fordern	4792	call (45.3%); urge (22.2%); ask (19.3%)	chiedere (40.4%); invitare (26.8%); esortare (19.6%)
zu	stimmen	4383	agree (80.3%)	concordare (51.7%); condividere (15.8%)
vor	schlagen	4200	propose (76.4%); suggest (17.6%)	proporre (85.2%)
fest	stellen	2538	be (31.9%); note (28.9%)	constare (16.8%)
vor	liegen	1891	be (45.9%); have (33.4%)	essere (28.0%); esserelsonare (17.8%)
vor	sehen	1873	provide (40.9%)	prevedere (72.1%)
überein	stimmen	1864	agree (78.1%)	concordare (55.1%); condividere (17.1%)
ab	lehnen	1690	reject (51.8%); oppose (19.3%)	respingere (40.8%); rifiutare (22.0%); opporre (15.2%)
aus	gehen	1617	be (32.5%); assume (20.3%)	
hin	weisen	1606	point (21.5%)	sottolineare (15.5%)
bei	tragen	1523	contribute (42.6%); help (20.5%)	contribuire (68.4%)
an	schließen	1425	agree (28.5%); join (16.5%)	associare (28.4%); condividere (19.8%); unire (17.6%)
aus	reichen	1361	be (63.9%); suffice (20.4%)	bastare (85.4%)
aus	sehen	1291	be (48.4%); look (18.4%)	essere (32.6%); sembrare (32.4%)
ab	zielen	1250	aim (56.8%)	mirare (25.9%)
ab	hängen	1202	depend (86.4%)	dipendere (89.0%)
an	nehmen	1086	adopt (27.1%); assume (17.1%)	
an	kommen	1057	be (71.3%)	essere (27.6%)
auf	rufen	1049	call (79.7%)	chiedere (35.9%); invitare (23.9%); esortare (18.0%)
aus	sprechen	1010	be (30.0%)	esprimere (15.9%)

(2) Übersetzungsvarianten von Mehrwortausdrücken

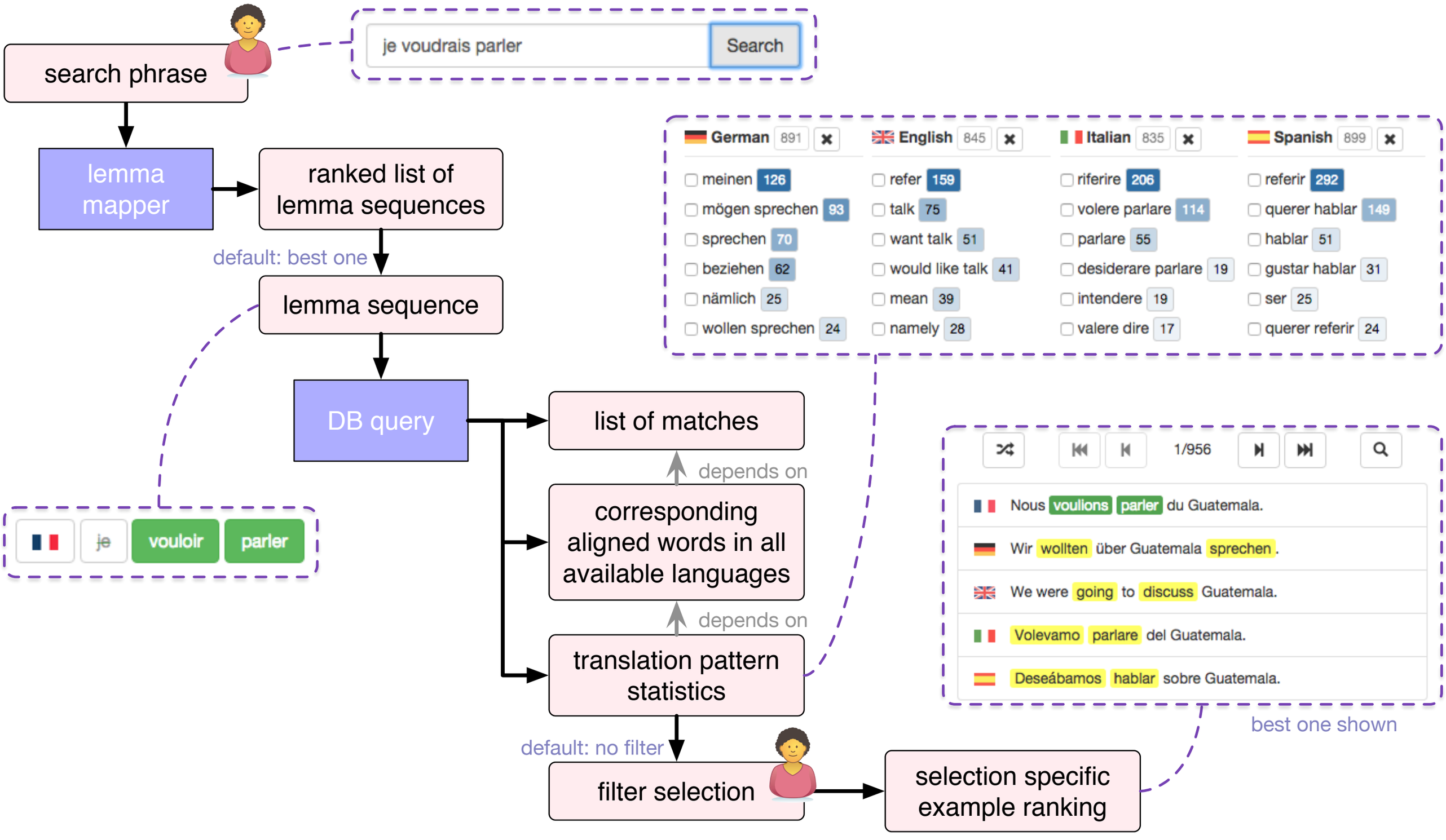
- *Multilingwis (Multilingual Word Information System)*

  - Webanwendung mit Suchfeld für beliebige Mehrwortausdrücke in (z.Zt.) Deutsch, Englisch, Französisch, Spanisch und Italienisch

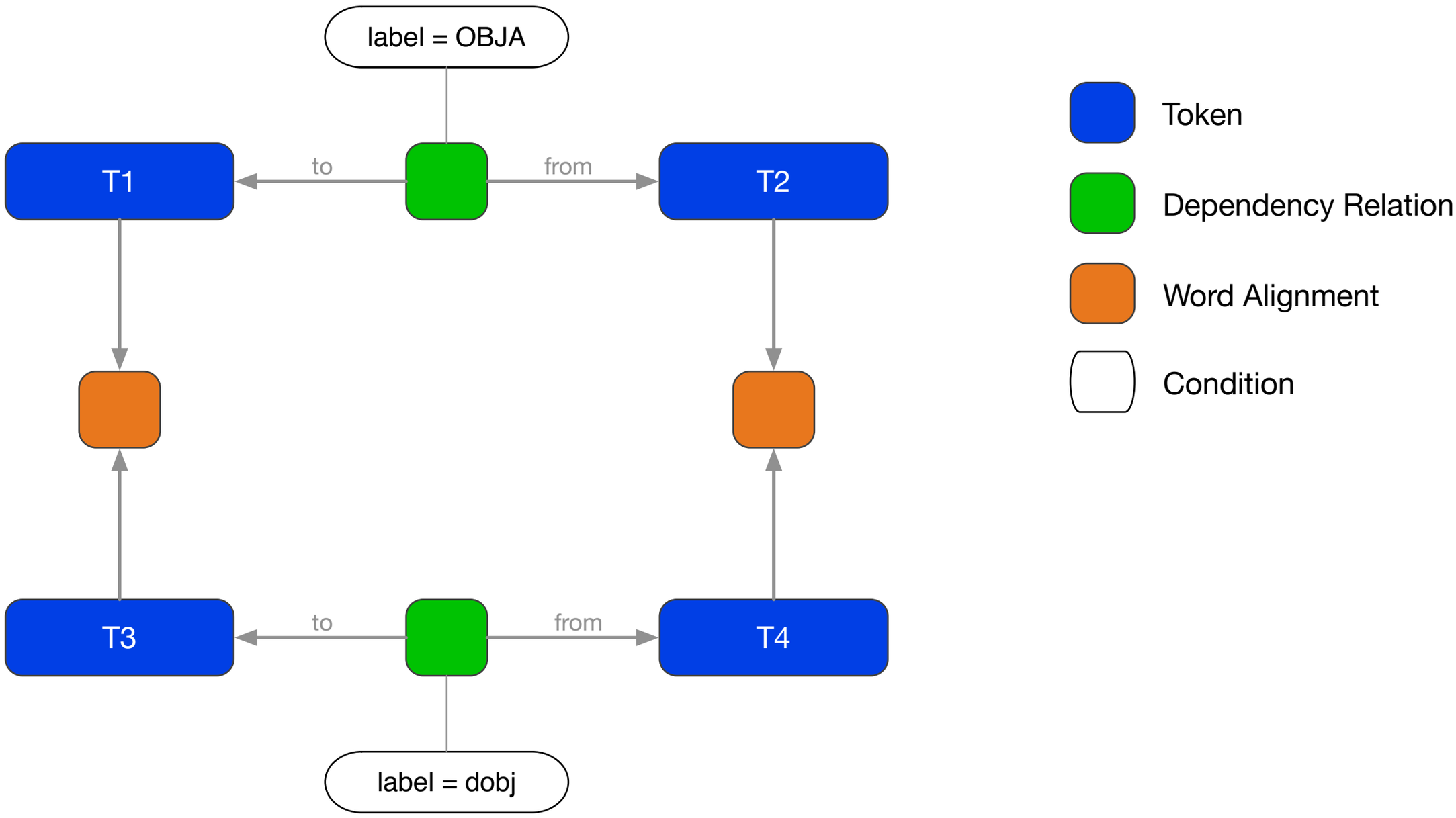
  - nach Häufigkeit sortierte Übersetzungsvarianten

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

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
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

(3) Funktionsverbgefüge anhand der Alignierung identifizieren



Ergebnisse: de/it

Zeilen 1..21 / 31896

T1	T2	T3	T4	f(T1..T4)	f(T1,T2)	f(T3,T4)	O/E	t-score
fordern	Tribut	avere	prezzo	2	3	48	0.026	-0.109
haben	Wort	concedere	parola	2	251	92	0.025	-0.055
haben	Verständnis	rendere	conto	5	45	15	0.047	-0.073
bieten	Möglichkeit	avere	possibilità	2	333	1062	0.036	-0.055
haben	Funktion	svolgere	funzione	3	39	133	0.089	-0.052
haben	Wahl	svolgere	elezione	2	51	23	0.089	-0.052
sein	Aufgabe	avere	dovere	2	2	218	0.380	-0.103
haben	Bedeutung	assumere	significato	2	145	12	0.079	-0.038
haben	Auswirkung	determinare	impatto	2	786	5	0.061	-0.025
haben	Wort	prendere	parola	6	251	238	0.142	-0.040
tun	Recht	avere	ragione	2	2	507	0.156	-0.043
tun	Leid	chiedere	scusa	2	2	2	0.076	-0.022
haben	Vorteil	comportare	vantaggio	7	99	44	0.149	-0.024
haben	Auswirkung	comportare	implicazione	2	786	4	0.149	-0.024
haben	Problem	vedere	problema	2	474	28	0.193	-0.028
spielen	Rolle	avere	importanza	3	2127	57	0.162	-0.020
spielen	Rolle	avere	ruolo	2	2127	72	0.162	-0.020
spielen	Rolle	avere	compito	2	2127	144	0.162	-0.020
sehen	Problem	avere	problema	2	43	432	0.269	-0.025
haben	Frage	formulare	domanda	2	176	51	0.143	-0.018
beziehen	Rente	avere	pensione	2	6	8	0.171	-0.019

	1	2	3	4	5	6	7	8
F	Auch	hier	fordert	die	Individualisierung	ihren	Tribut	.
L	auch	hier	fordern	die	Individualisierung	ihr	Tribut	.
T	ADV	ADV	VVFIN	ART	NN	PPOSAT	NN	$.

	1	2	3	4	5	6	7	8	9	10
F	Anche	in	questo	caso	l’	individualizzazione	ha	un	prezzo	.
L	anche	in	questo	caso	il	individualizzazione	avere	un	prezzo	.
T	ADV	PRE	PRO:demo	NOM	DET:def	NOM	VER:pres	DET:indef	NOM	SENT

	1	2	3	4	5	6	7
F	Dieser	Kundendienst	fordert	jedoch	seinen	Tribut	.
L	dies	Kundendienst	fordern	jedoch	sein	Tribut	.
T	PDAT	NN	VVFIN	ADV	PPOSAT	NN	$.

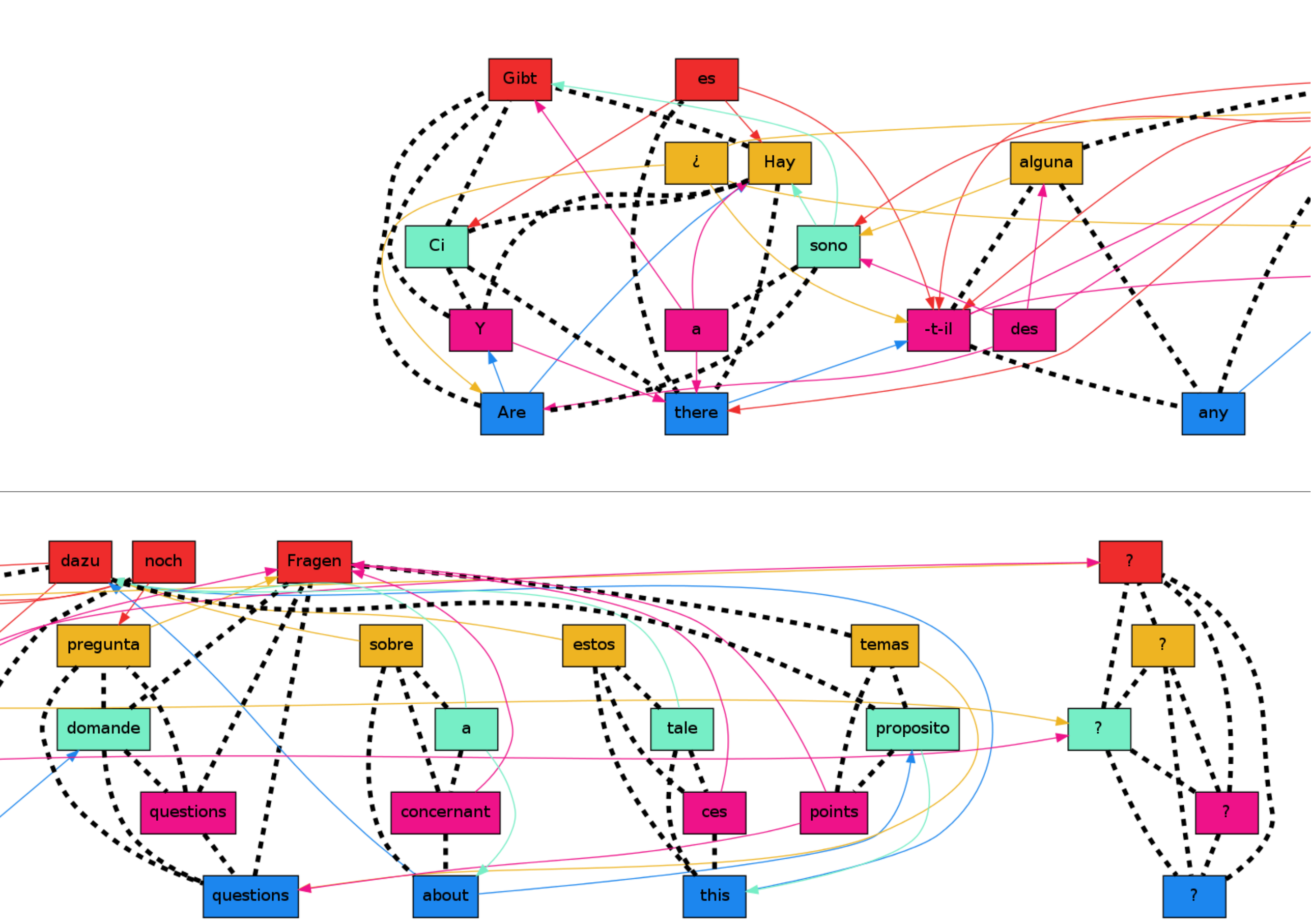
	1	2	3	4	5	6	7	8	9	10	11
F	Le	relazioni	commerciali	di	questo	tipo	hanno	però	un	prezzo	.
L	il	relazione	commerciale	di	questo	tipo	avere	però	un	prezzo	.
T	DET:def	NOM	ADJ	PRE	PRO:demo	NOM	VER:pres	ADV	DET:indef	NOM	SENT

Ergebnisse: en/it

Zeilen 1..21 / 57781

T1	T2	T3	T4	f(T1..T4)	f(T1,T2)	f(T3,T4)	O/E	t-score
have	directive	approvare	direttiva	2	79	82	0.003	-0.294
have	debate	sviluppare	dibattito	2	299	5	0.006	-0.141
represent	volume	avere	volume	2	3	8	0.008	-0.115
have	amendment	proporre	emendamento	2	40	256	0.014	-0.158
have	problem	affrontare	problema	3	873	1275	0.014	-0.149
have	difficulty	affrontare	difficoltà	2	358	51	0.014	-0.149
know	border	avere	confine|confino	2	25	25	0.015	-0.146
know	bound	avere	limite	2	13	66	0.015	-0.146
have	result	comunicare	risultato	2	75	21	0.010	-0.068
have	value	rappresentare	valore	2	179	40	0.023	-0.125
have	county	comprendere	contea	2	2	2	0.023	-0.093
have	agreement	concludere	accordo	3	141	364	0.021	-0.082
need	will	avere	volontà	2	6	90	0.043	-0.093
have	definition	fornire	definizione	2	19	24	0.033	-0.082
have	view	fornire	visione	2	149	15	0.033	-0.082
have	tool	fornire	strumento	2	42	83	0.033	-0.082
have	hearing	organizzare	audizione	2	13	57	0.021	-0.056
have	right	riconoscere	diritto	6	1967	291	0.034	-0.074
call	shot	fare	bello	2	2	3	0.024	-0.050
have	policy	attuare	politica|politico	5	299	311	0.032	-0.059
have	policy	adottare	politica|politico	7	299	137	0.062	-0.071

Ausblick


Gibt
es
¿
Hay
alguna
Ci
sono
Y
a
-t-il
des
Are
there
any
dazu
noch
Fragen
?
pregunta
sobre
estos
temas
?
?
domande
a
tale
proposito
?
questions
concernant
ces
points
?
questions
about
this
?