InterCorp: viele Sprachen – ein Korpus

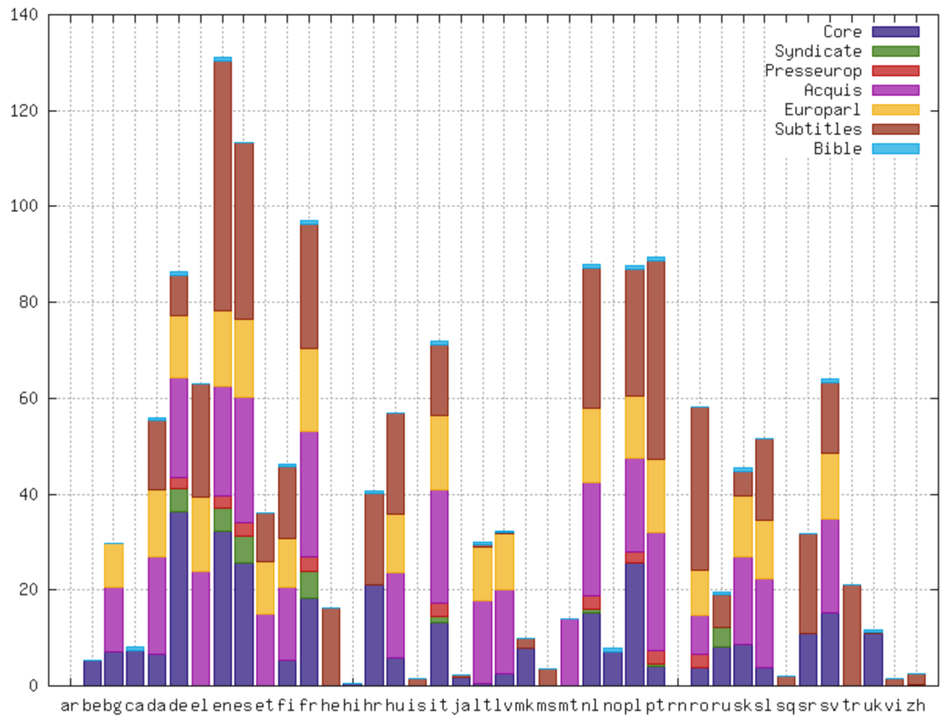
Ein multilinguales Parallelkorpus (nicht nur) europäischer Sprachen

CZECH NATIONAL CORPUS

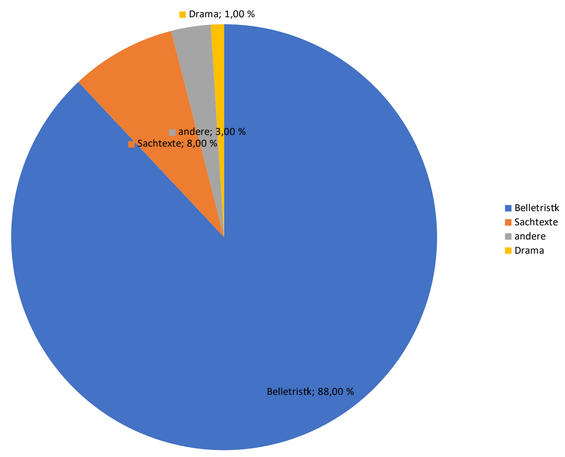
InterCorp entsteht seit 2005 am Institut des Tschechischen Nationalkorpus, Philosophische Fakultät, Karls-Universität (Prag, Tschechien). Die erste Ausgabe ging am 19. 11. 2009 online und beinhaltete 22 Sprachen (letzte Version 12 vom 12. 12. 2019). Am Aufbau des Kernkorpus (manuelle Aufbereitung der Texte) beteiligen sich auch Studierende der jeweiligen Sprache. Der Zugang zur Vollversion ist kostenlos und für akademische Zwecke bestimmt.

Aufteilung der Texte

sog. Kollektionen: Core (Kernkorpus), Syndicate, Presseurop, Acquis, Europarl, Subtitles, Bible

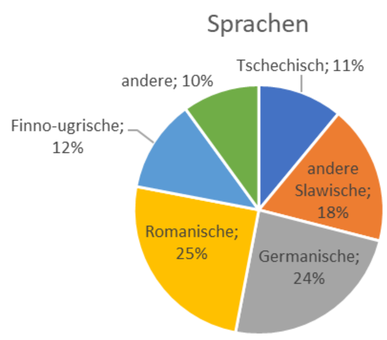


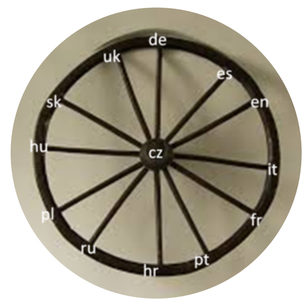
Stilistik im Kernkorpus (Core):



Grunsbeschreibung

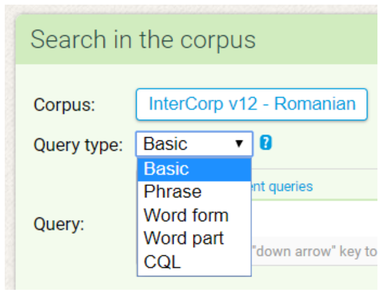
Paralelle alignierte Texte
Tschechisch und anderssprachige Versionen





Morphosyntaktische Annotation			
Sprache	**Tagging**	**Lemmat.**	**Tool**
Albanisch			
Arabisch			
Bulgarisch	✔	✔	TreeTagger
Chinesisch	✔		ZPar v0.7.5
Dänisch			
Deutsch	✔	✔	RFTagger
Englisch	✔	✔	TreeTagger
Estnisch	✔	✔	TreeTagger
Finnisch	✔	✔	OMorFi +HunPOS
Französisch	✔	✔	TreeTagger
Griechisch			
Hebräisch			
Hindu			
Isländisch	✔	✔	IceStagger
Italienisch	✔	✔	TreeTagger
Japanisch	✔	✔	MeCab + Unidic
Japanisch			
Katalanisch	✔	✔	TreeTagger
Kroatisch	✔	✔	ReLDI Tagger
Lettisch	✔	✔	LVTagger
Litauisch			
Malaysisch			
Maltesisch			
Mazedonisch			
Niederländisch	✔	✔	TreeTagger
Norwegisch	✔	✔	VISL
Polnisch	✔	✔	Morfeusz, KRNNT
Portugiesisch	✔	✔	TreeTagger
Romanes			
Rumänisch			
Russisch	✔	✔	TreeTagger
Schwedisch	✔	✔	Stagger
Serbisch	✔	✔	ReLDI Tagger
Slowakisch	✔	✔	Radovan Garabík, Morče
Slowenisch	✔	✔	ToTaLe
Spanisch	✔	✔	TreeTagger
Tschechisch	✔	✔	Morče
Türkisch			
Ukrainisch	✔	✔	UDPipe
Ungarisch	✔		RFTagger
Vietnamesisch			
Weißrussisch	✔	✔	UDPipe

Abfragemöglichkeiten



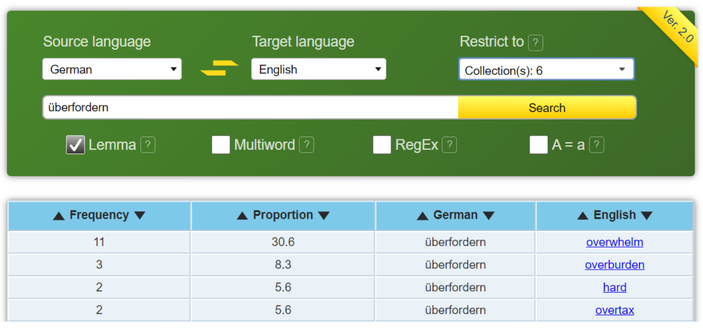
Beispiel: Abfragemodus

Weitere Funktionen (Auswahl):

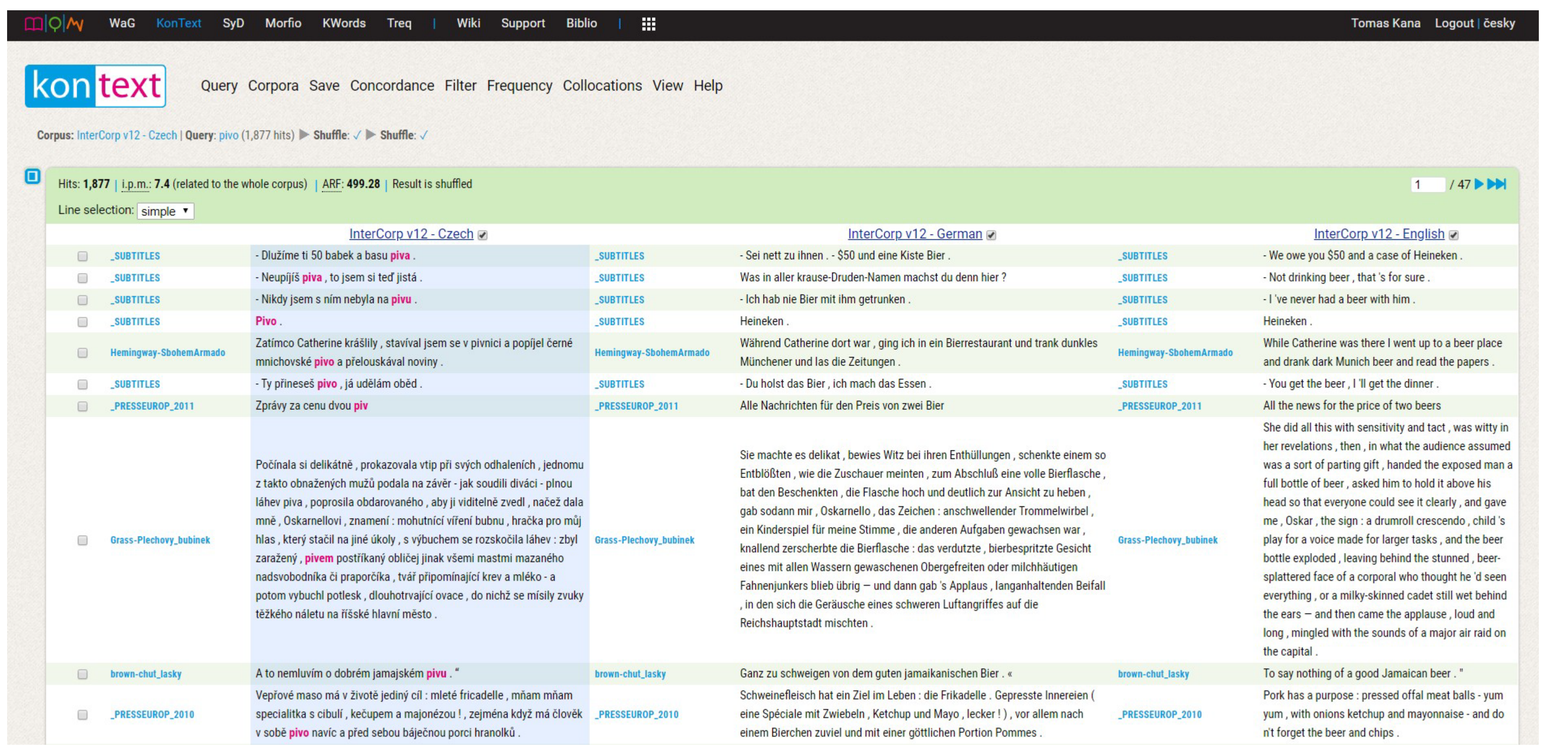
- Sortieren von Konkordanzen
- Paralelle Suche
- Filter
- Frequenzen (Formen, Texte, KWIC-Umfeld)
- Kollokationen

"Nebenprodukt"

Treq TRANSLATION EQUIVALENTS DATABASE



▲ Frequency ▼	▲ Proportion ▼	▲ German ▼	▲ English ▼
11	30.6	überfordern	overwhelm
3	8.3	überfordern	overburden
2	5.6	überfordern	hard
2	5.6	überfordern	overtax



Zugriffe	
Land	**Anzahl (IP)**
Tschechien	168 295
Deutschland	9 779
Polen	3 610
Bulgarien	3 080
Großbritannie	3 027
Rußland	2 711
Spanien	2 681
Italien	2 107
Slowakei	2 097
Frankreich	1 818

CZECH NATIONAL CORPUS
https://korpus.cz/
Tomáš Káňa, Martin Vavřín, Michal Křen
kana@ped.muni.cz