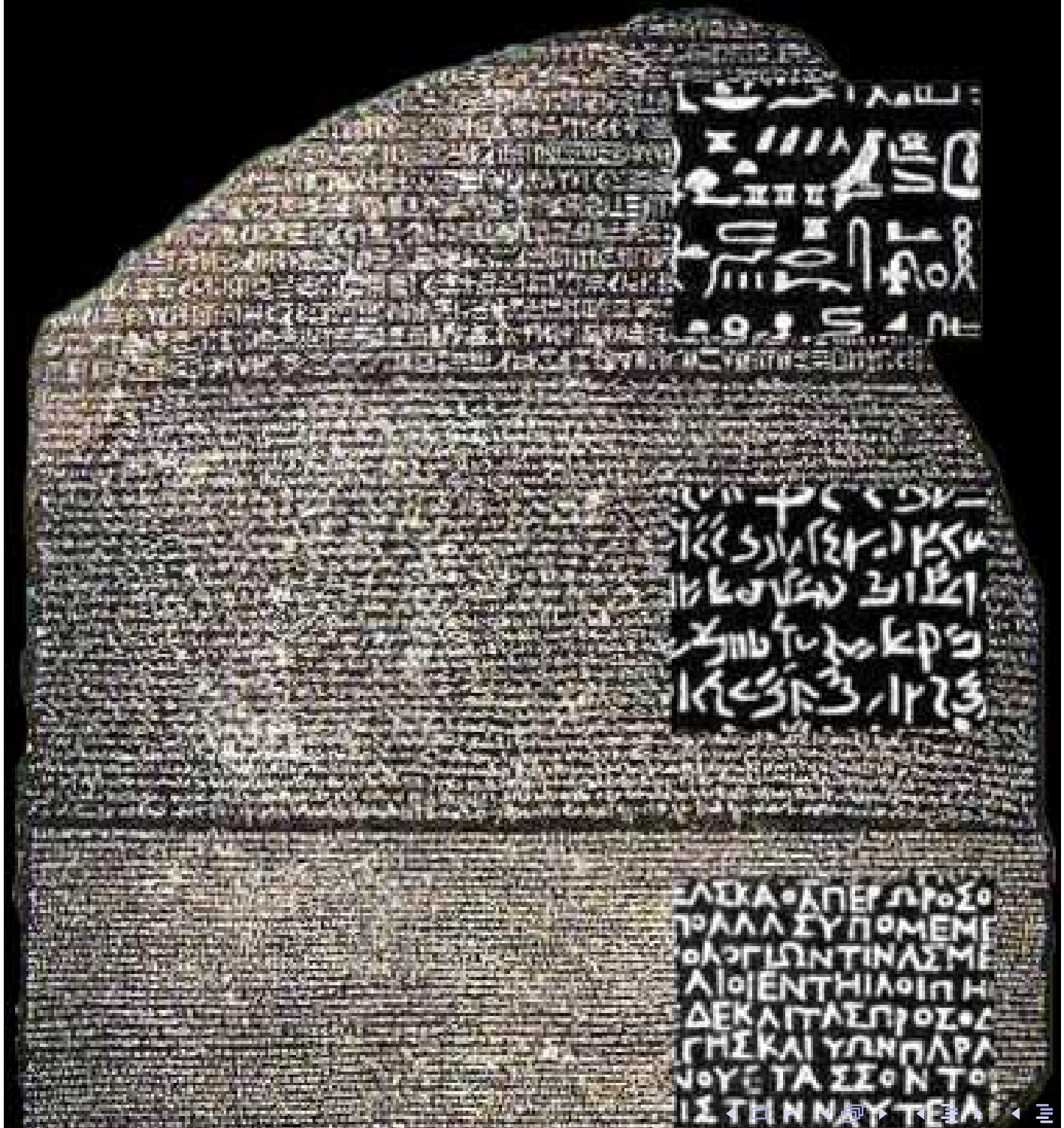


InterCorp: Exploring a multilingual parallel corpus

Alexandr Rosen

Institute of Theoretical and Computational Linguistics
Institute of the Czech National Corpus
Charles University in Prague, Faculty of Arts

Grammar and Corpora 2016
Institut für Deutsche Sprache, Mannheim
8–11 November 2016

Outline

Outline

- Same text in multiple versions (languages, translations, ...)
- Parallel vs. comparable
- Alignment by text units
- Problems:
  - authenticity
  - availability
  - alignment
  - tools
- Useful for:
  - translators (CAT tools)
  - FLT
  - lexicographers
  - researchers
  - NLP (machine translation, information retrieval, projecting annotation)

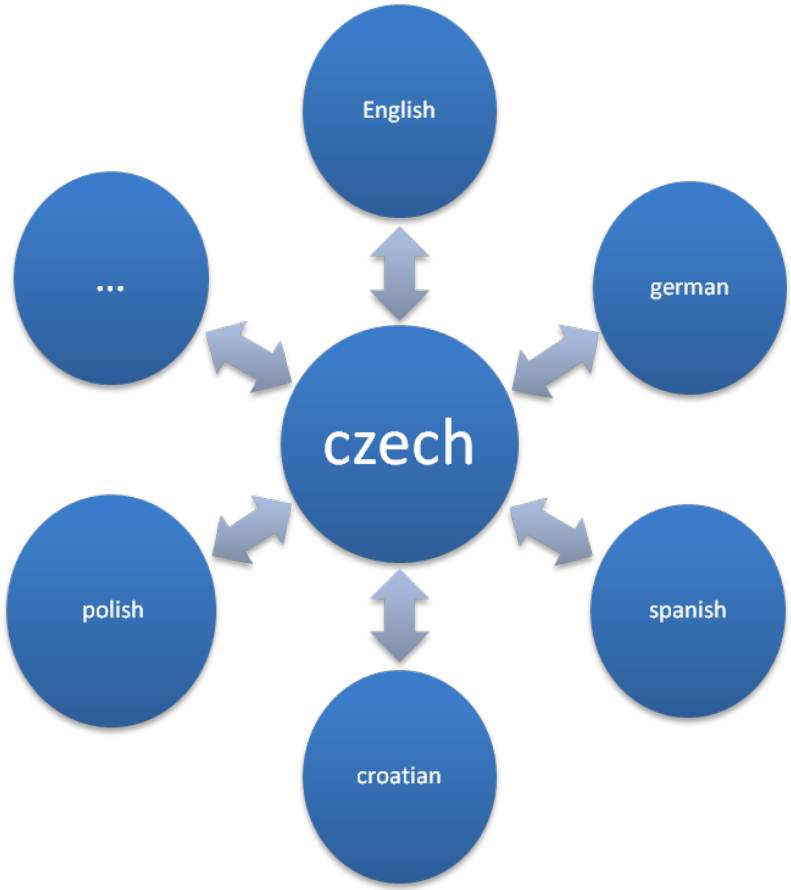
Outline

Basics

- A part of the *Czech National Corpus*
- Czech as the pivot language
- http://www.korpus.cz/intercorp/
- *2005, as a service for the linguistic departments of Charles University's Faculty of Arts
- Now used widely within and beyond the academia
- New releases once per year

The architecture of *InterCorp*

- Alignment: sentence-level, stand-off
- Each text in Czech and at least one other language
- Texts in other languages aligned via Czech
- Morphological tags and lemmatization – where the tools are available



39 languages + Czech

- 10 Slavic: be, bg, hr, mk, pl, ru, sk, sl, sr, uk
- 7 Germanic: da, de, en, is, nl, no, sv
- 6 Romance: ca, es, fr, it, pt, ro
- 16 other: ar, el, et, fi, he, hi, hu, ja, lt, lv, ms, mt, rn, sq, tr, vi

☛ Only few texts are available in more than 20 languages

☛ Languages differ wildly in the volumes of text

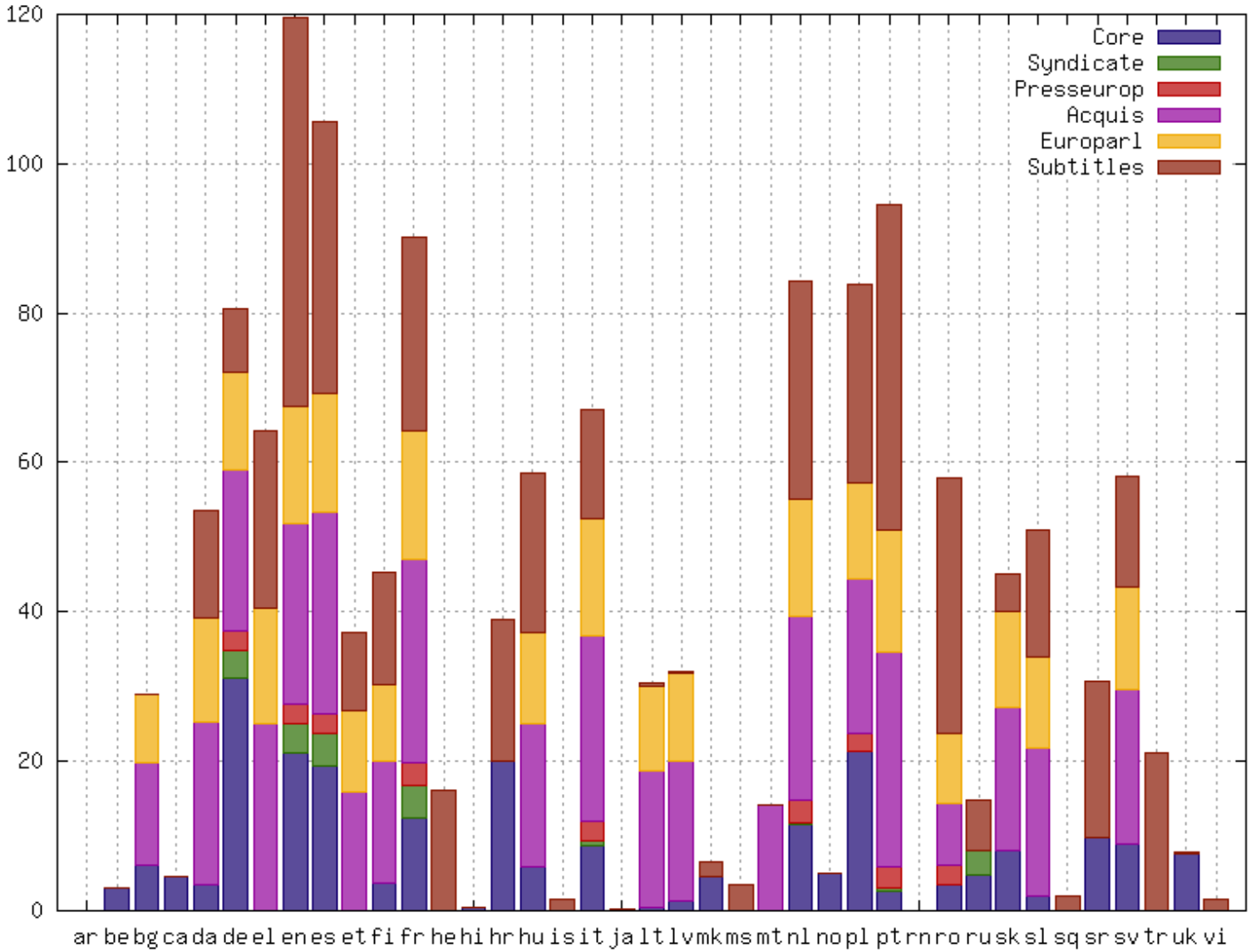
Text types

- **Total size** – 1.4 billion words; the Czech part: 187 million words
- The **core** – mostly fiction with manually checked alignment
- **Collections** – freely available texts with automatic alignment
  - **Journalism**
    Project Syndicate http://www.project-syndicate.org/
    VoxEurope http://www.voxeurop.eu/
  - **Law**
    Acquis Communautaire
    http://langtech.jrc.ec.europa.eu/JRC-Acquis.html
  - **Parliament proceedings**
    Europarl http://www.statmt.org/europarl/
  - **Film subtitles**
    Open Subtitles http://www.opensubtitles.org

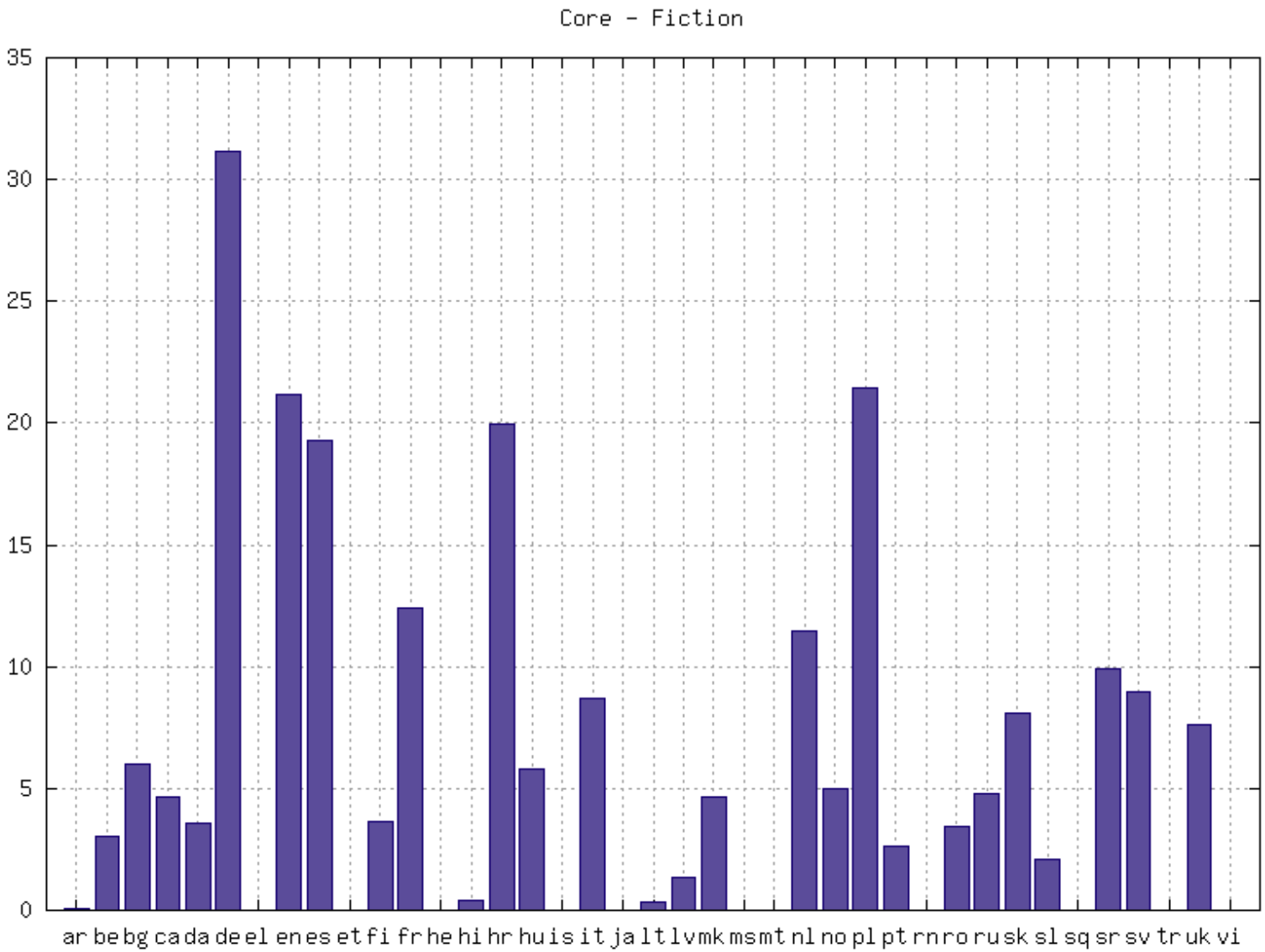
Size in million words

	Czech	German	All foreign	**Total**
Core	97.0	31.2	231.5	**328.5**
Syndicate	3.4	3.7	20.8	**24.2**
VoxEurope	2.3	2.5	24.7	**27.0**
Acquis	20.3	21.7	430.2	**450.5**
Europarl	12.9	13.1	277.9	**290.8**
Subtitles	50.7	8.4	488.3	**539.0**
Total	186.6	80.6	1473.4	**1660.0**
No. of core texts	1,435	362	3,024	**4,459**

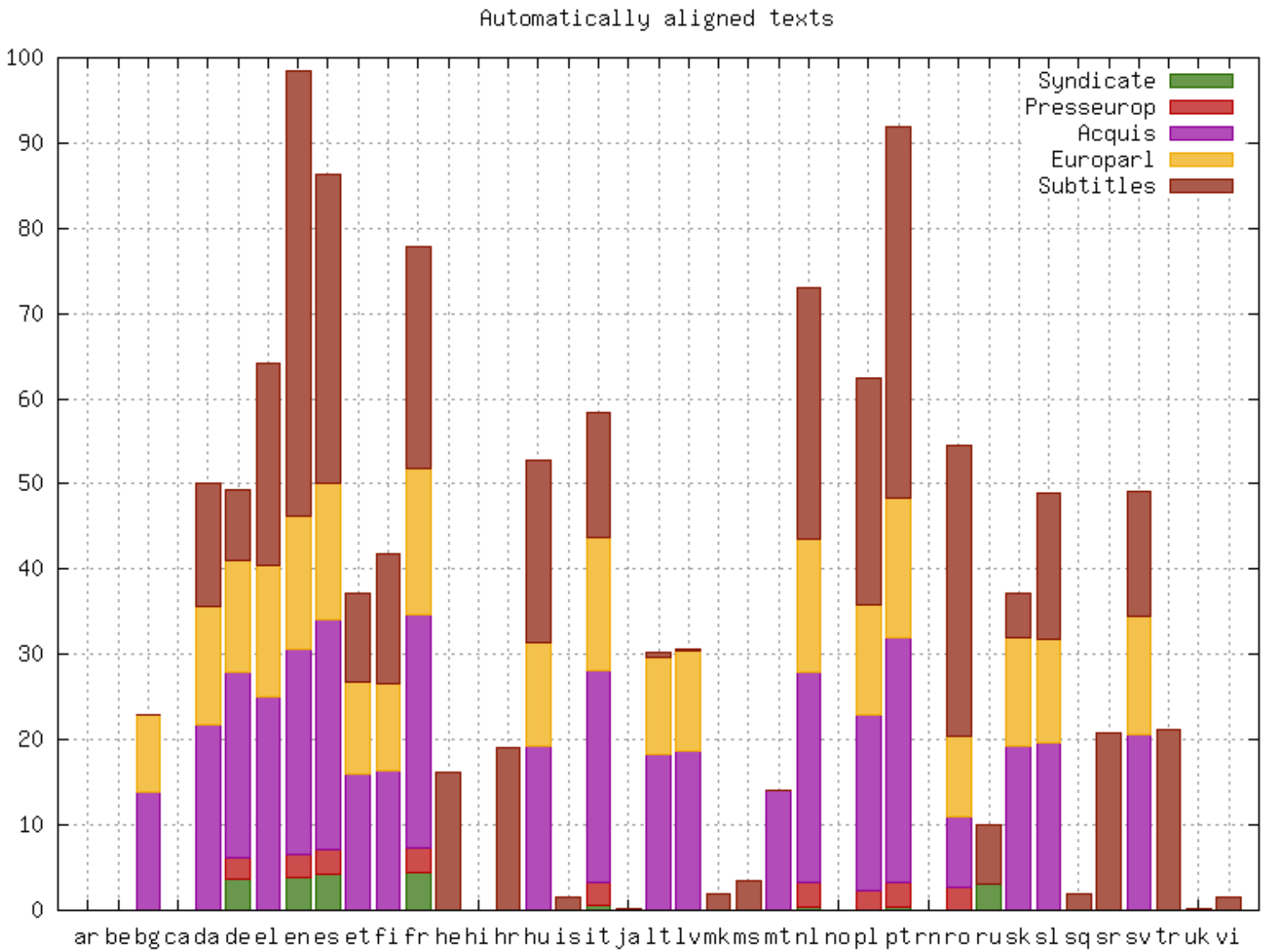
InterCorp by languages and text types



Core (fiction)



Collections (journalism, law, parliament proceedings)



No. of languages	Author	Title
26	Rowling	*Harry Potter and the Philosopher's Stone*
26	Saint-Exupéry	*The Little Prince*
23	Carroll	*Alice in Wonderland*
21	Kundera	*The Unbearable Lightness of Being*
21	Rowling	*Harry Potter and the Chamber of Secrets*
21	Tolkien	*The Fellowship of the Ring*
20	Kundera	*The Joke*
20	Adams	*The Hitch Hiker's Guide to the Galaxy*
20	Tolkien	*The Return of the King*
19	Bulgakov	*The Master and Margarita*
19	Rowling	*Harry Potter and the Prisoner of Azkaban*
19	Brown	*The Da Vinci Code*
19	Tolkien	*The Two Towers*
18	Tolkien	*The Hobbit or There and Back Again*
18	Hašek	*The Good Soldier Švejk*
18	Eco	*The Name of Rose*
18	Milne	*Winnie the Pooh*
17	Orwell	*1984*
17	Kafka	*The Trial*
17	Rowling	*Harry Potter and the Goblet of Fire*
17	Coelho	*The Alchemist*
16	Kundera	*Immortality*
16	Frank	*The Diary of a Young Girl*
16	Hrabal	*I Served the King of England*
16	Kipling	*The Jungle Book*
15	Kundera	*Laughable Loves*
15	Rowling	*Harry Potter and the Order of the Phoenix*

Outline

Name	Types	Langs	Size	Annot	Aligned	Proofread	Search	Download	Metadata
Linguee	legal	25	?	no	S,W	no	yes	no	yes
Glosbe	varia	100+	1Bs	no	S,W	no	yes	no	yes
SKE	varia	38	217M cs	no	S	no	yes	yes	yes
DGT-TM	legal	22	3.7Ms cs	no	S	yes	no	yes	no
Pelcra	varia	31	58M pl	no	S,W	part	no	yes	yes
RNC	varia	6	9M	M	S	part	yes	?	yes
SNK	fiction	7	388M sk	M	S	no	yes	part	yes
CzEng	varia	en, cs	233M en	M,Sy	S	no	sample	yes	no
PCEDT	news	en, cs	1.2M.	M,Sy,Se	S,W	yes	yes	yes	yes
Kačenka	fiction	en, cs	3.3M	no	S	yes	no	yes	yes
Opus	varia	100+	4.7B	M,Sy	S,W	no	yes	yes	no
Parasol	fiction	31	27M	M	S	part	yes	?	yes
ASPAC	fiction	25	68t	no	P	yes	no	?	yes
InterCorp	varia	32	1.6B	M	S	part	yes	yes	yes

- **Linguee**: online search through bilingual texts – http://www.linguee.com
- **Glosbe**: Translation Memory Online – http://glosbe.com/tmem/
- **SKE**: Sketch Engine – http://www.sketchengine.co.uk
- **DGT-TM**: Translation Memory of the European Commission's Directorate-General for Translation – http://ipsc.jrc.ec.europa.eu/?id=197
- **Pelcra**: Polish & English Language Corpora for Research & Applications – http://pelcra.pl/new/
- **RNC**: Russian National Corpus – http://www.ruscorpora.ru
- **SNK**: Slovak National Corpus – http://korpus.juls.savba.sk/par.html
- **CzEng**: Czech-English parallel corpus – http://ufal.mff.cuni.cz/czeng
- **PCEDT**: Prague Czech-English Dependency Treebank – http://ufal.mff.cuni.cz/prague-czech-english-dependency-treebank
- **Kačenka**: English-Czech Corpus of the Department of English Studies, Faculty of Arts, Masaryk University Brno – http://www.phil.muni.cz/angl/kacenka/kachna.html

Outline

On-line queries

KonText

http://kontext.korpus.cz

- Choice of *InterCorp* release
- Search filters:
  - languages, texts, publication year, text type
  - original/translation, language of the original
  - sex of the author and the translator
- Parallel queries, CQL
- Positive and negative filters on the concordances
- Export of concordances
- Sorting, frequency distribution, collocations
- Custom subcorpora

Treq – translation equivalents

- treq.korpus.cz
- Based on word-to-word alignment

German equivalents for Czech *křičet*

1135	*schreien*	2	*losschreien*
185	*rufen*	2	*Rufen*
60	*brüllen*	1	*aufheulen*
31	*anschreien*	1	*dazwischenrief*
21	*Schrei*	1	*durcheinanderschrien*
16	*schreiend*	1	*greinte*
10	*kreischen*	1	*grölen*
6	*aufschreien*	1	*herriefen*
6	*Geschrei*	1	*herumzubrüllen*
3	*zurufen*	1	*hinausschreien*
3	*zuschrie*	1	*Lamentieren*
2	*anschreie*	1	*nachrufen*
2	*brüllend*	1	*quieken*
2	*geschrieen*	1	*schreit*
2	*herumschreien*	1	*schriest*
2	*hineinrufen*	1	*stöhnen*

Dissemination of texts

- Technical protection against misuse:
  shuffled order of blocks of translation pairs
- Blocks of sentence in pairs up to 100 words as units
- Only 1:1 alignment pairs
- Educational and research licence, no re-distribution

Outline

Pre-processing

1. Acquisition
2. Scanning & character recognition
3. Proofreading
4. Segmentation (sentence boundary detection)
5. Alignment
6. Checking of segmentation and alignment
7. Morphosyntactic markup

Tools used in pre-processing

1. Bibliographical database
2. *Intertext* – alignment editor
3. *Punkt* – sentence splitter
4. *Hunalign* – aligner
5. Language-specific tokenizers and taggers

Linguistic markup

= lemmatization and tagging
by morphosyntactic and morphological categories

Strategy

- Use available tools (taggers), including:
  - Tokenization bundled with the tool
  - Tagsets designed elsewhere by experts on the given language
  - Annotation models trained elsewhere

Current state

- Tags for Czech + 23 foreign languages
- Lemmas for Czech + 20 foreign languages

```
<?xml version='1.0' encoding='utf-8'?>
<!DOCTYPE doc SYSTEM https://trnka.ff.cuni.cz/ucnk/intercorp/files/intercorp.dtd>
<doc id="Hasek-OsudyDobrehoVvSV" language="cs" version="00">
<p id="cs:Hasek-OsudyDobrehoVvSV:0:1">
<s id="cs:Hasek-OsudyDobrehoVvSV:0:1:1">
<w lemma="1" tag="ClFS1-----------">1</w>
</s>
</p>
<p id="cs:Hasek-OsudyDobrehoVvSV:0:2">
<s id="cs:Hasek-OsudyDobrehoVvSV:0:2:1">
<w lemma="zasáhnutí" tag="NNNS1-----A-----">ZASÁHNUTÍ</w>
<w lemma="dobrý" tag="AAMS2----1A-----">DOBRÉHO</w>
<w lemma="voják" tag="NNMS2-----A-----">VOJÁKA</w>
<w lemma="švejk" tag="NNMS2-----A-----">ŠVEJKA</w>
<w lemma="do" tag="RR--2-----------">DO</w>
<w lemma="světový" tag="AAFS2----1A-----">SVĚTOVÉ</w>
<w lemma="válka" tag="NNFS2-----A-----">VÁLKY</w>
</s>
</p>
<p id="cs:Hasek-OsudyDobrehoVvSV:0:3">
<s id="cs:Hasek-OsudyDobrehoVvSV:0:3:1">
<w lemma="&amp;quot;" tag="Z:--------------">&amp;quot;</w>
<w lemma="tak" tag="Db-------------">Tak</w>
<w lemma="já" tag="PP--3--1--------">nám</w>
<w lemma="zabít" tag="VpMP---3R-AA---P">zabili</w>
<w lemma="ferdinand" tag="NNMS4-----A-----">Ferdinanda</w>
<w lemma="," tag="Z:--------------">,</w>
<w lemma="&amp;quot;" tag="Z:--------------">&amp;quot;</w>
<w lemma="říci" tag="VpFS---3R-AA---B">řekla</w>
```

```
<p id="pl:Hasek-OsudyDobrehoVvSV:0:12">
<s id="pl:Hasek-OsudyDobrehoVvSV:0:12:1">
<w lemma="jak" tag="conj" conf="dd">JAK</w>
<w lemma="dobry" tag="adj:sg:nom:m1:pos" conf="dd">DOBRY</w>
<w lemma="wojak" tag="subst:sg:nom:m1" conf="dd">WOJAK</w>
<w lemma="szwejk" tag="subst:sg:nom:m1" conf="dd">SZWEJK</w>
<w lemma="wkroczyć" tag="praet:sg:m1:perf" conf="dd">WKROCZYŁ</w>
<w lemma="na" tag="prep:acc" conf="dd">NA</w>
<w lemma="widownia" tag="subst:sg:acc:f" conf="dd">WIDOWNIĘ</w>
<w lemma="wojna" tag="subst:sg:gen:f" conf="dd">WOJNY</w>
<w lemma="światowy" tag="adj:sg:gen:f:pos" conf="dd">ŚWIATOWEJ</w>
</s>
</p>
<p id="pl:Hasek-OsudyDobrehoVvSV:0:13">
<s id="pl:Hasek-OsudyDobrehoVvSV:0:13:1">
<w lemma="-" tag="interp" conf="dd">-</w>
<w lemma="a" tag="conj" conf="dd">A</w>
<w lemma="to" tag="subst:sg:acc:n" conf="dd">to</w>
<w lemma="my" tag="ppron12:pl:dat:f:pri" conf="dd">nam</w>
<w lemma="zabić" tag="praet:pl:m1:perf" conf="dd">zabili</w>
<w lemma="ferdynand" tag="subst:sg:acc:m1" conf="dd">Ferdynanda</w>
<w lemma="-" tag="interp" conf="dd">-</w>
<w lemma="rzec" tag="praet:sg:f:imperf" conf="dd">rzekła</w>
<w lemma="posługaczka" tag="subst:sg:nom:f" conf="dd">posługaczka</w>
<w lemma="do" tag="prep:gen" conf="dd">do</w>
<w lemma="pan" tag="subst:sg:gen:m1" conf="dd">pana</w>
<w lemma="szwejk" tag="subst:sg:gen:m1" conf="dd">Szwejka</w>
<D/>
<w lemma="," tag="interp" conf="dd">,</w>
```

Tools used for lemmatization and tagging

Lng	Tags	Lms	Tool	Preposition Determiner Adjective Noun
bg	✓		TT	R Pde-os-n Ansi Ncnsi
cs	✓	✓	Morče	RR-6 PDXP6 AAFP6---3A NNFP6---A
de	✓	✓	TT	APPR ART ADJA NN
en	✓	✓	TT	IN DT JJS NNS
es	✓	✓	TT	PREP ART NC ADJ
et	✓	✓	TT	P--s3 A-p-s3 Nc-s3
fr	✓	✓	TT	PRP DET:ART ADJ NOM
hu	✓		HunPos	ART ADJ ADJ NOUN (CAS (ILL))
it	✓	✓	TT	PRE PRO:demo NOM ADJ
lt	✓	✓	V.D.	prln jvrd bdvr dktv
nl	✓		TT	600 370 103 000
no	✓	✓	OB	prep det adj subst
pl	✓	✓	TaKIPI	prep:loc:nwok adj:sg:loc:m3:pos adj:sg:loc:m3:pos subst:sg:loc:m3
pt	✓	✓	TT	SPS DA0 NCFS AQ0
ru	✓	✓	TT	Sp-l P--pl Afp-plf Ncmpln
sk	✓	✓	Morče	Eu6 PFfs6 AAfs6x SSfs6
sl	✓	✓	totale	Sl Pd-nsg Agpfsg Ncnsl

Outline

Access statistics

- January–June 2015
- For each user query:
  - filters for corpus parts (core, collections)
  - combination of languages
- Custom corpora not included
- Total no. of queries: 62 thousand $\approx$ 310 per day

No. of languages in a query

Languages	Queries	% queries
1	10 431	16.80 %
2	49 905	80.37 %
3	1 314	2.12 %
4	197	0.32 %
5	146	0.24 %
6	82	0.13 %
33	14	0.02 %
40	2	0.00 %
Σ	**62 091**	**100.00 %**

Most frequent language combinations

cs	en	17 504
cs	de	6 805
cs	es	6 239
cs	fr	3 484
cs	nl	2 446
es		2 144
cs		1 901
de		1 587
cs	pl	1 272
cs	it	1 213
cs	fi	1 165
en		1 037
cs	ru	1 035

Languages in queries

cs	50 619	ar	130
en	22 380	tr	118
de	10 542	uk	113
es	9 497	sr	106
fr	5 597	sl	67
nl	2 949	mk	55
fi	2 162	ja	42
ru	2 147	no	22
it	2 126	da	14
pl	1 922	is	14
sv	780	be	11
sk	670	vi	11
bg	428	ca	10
lv	412	hi	8
el	373	ro	7
hr	286	sq	6
pt	258	he	5
hu	147	et	1
lt	138	Σ	**114 181**

Corpus parts

all	56,487	90.97%
Core	3,511	5.65%
Subtitles	622	1.00%
Syndicate	522	0.84%
Europarl	311	0.50%
Acquis	296	0.48%
VoxEurop	142	0.23%
collections	109	0.18%
Acquis|Europarl	91	0.15%
Σ	**62,091**	**100.00%**

Queries / size ratio by language

	Total	Queries	**Core**	Queries		**Total**	Queries	**Core**	Queries
ar	52,99	130	7,60	6	ja	5,19	42		0
be	0,07	11	0,02	1	lt	0,06	138	0,00	0
bg	0,21	428	0,07	9	lv	0,18	412	0,03	1
ca	0,03	10	0,00	0	mk	0,14	55	0,06	5
cs	4,06	50 619	1,47	2 871	nl	0,50	2 949	1,80	412
da	0,00	14	0,06	4	no	0,06	22	0,01	1
de	1,91	10 542	0,17	106	pl	0,34	1 922	0,48	192
el	0,08	373		0	pt	0,04	258	0,22	12
en	2,75	22 380	3,93	1 402	ro	0,00	7	0,00	0
es	1,28	9 497	1,09	440	ru	2,24	2 147	0,68	52
et	0,00	1		0	sk	0,21	670	0,12	21
fi	0,67	2 162	2,26	178	sl	0,02	67	0,05	1
fr	0,90	5 597	1,86	392	sq	0,04	6		0
he	0,00	5		0	sr	0,05	106	0,07	14
hi	0,27	8	0,00	0	sv	0,19	780	0,04	8
hr	0,12	286	0,06	22	tr	0,08	118		0
hu	0,04	147	0,02	2	uk	0,30	113	0,03	4
is	0,12	14		0	vi	0,10	11		0
it	0,45	2 126	1,56	260	**Ø/Σ**	**1,78**	**114 181**	**0,72**	**6 416**

Survey (winter 2015/2016)

- **Asked** 748 users, response rate 17.4% (130)
- Most popular **conbination** with cs: en, en-de, de, en-fr, es, fr
- Most popular **text types**: core (62%), whole corpus (42%), Syndicate (17%), Europarl (14%), Acquis (13%), VoxEurop (12%), Subtitles (8%)
- **Motivation**: contrastive analysis (74%), equivalents (69%), translatological analysis (61%), single language (27%), FLT preparation (22%), FLT in-class (21%), NLP (4%)
- **Desiderata**: larger core (55 %), more translations in one language (52%), other text types (42%), inclusion of the original (41%), as many translations of a text as possible (31%), tagset harmonization (30%), balanced subcorpora (25%), other languages (7%)
- **Positives**: choice of languages, correct alignment, partitioning and custom corpora, extensions and improvements, free access, user support, availability
- **Negatives**: size, translation quality (Subtitles), disparate tagets, missing metadata, technica drawbacks of the interface

Outline

Content

- More **representative/balanced** core
  genres, periods, originals/translations, authors, translators
  – needed for both contrastive and translatological studies
- **The more the better**
  – the overlap may be too small even for languages such as English or German
- The **original text** should always be included
- **Multiple translations** in a single language

Search interface

- Missing functionalities:
  - **biKWiC** – highlighting keyword equivalent
  - Info on **alignment**: 1:1 / 2:1 / 1:2 / automatic / manual / confidence score
  - **Labeling**/annotating concordances
- Tools beyond mere search
  - **Comparisons** across text types, languages, corpora ...
  - **Co-occurence** profiles [Belica(2011)]
  - Word **sketches** [Kilgarriff et al.(2014)]

Annotation

- Languages differ in tagsets and tokenization rules
  – **harmonization of tagsets**
- Morphosyntactic annotation **for all languages**
- **Alignment** by words, multiword units, constituents
- **Syntactic** annotation
- **Crowdsourcing** to eliminate annotation errors

- *abychom, udělals, tys, očs, zum, aux*
  ×
  že by śmy, zrobił eś, ty ś, gdzieś/gdzie ś, ca n't, I 'm
- *Estados~Unidos* (NP), *a~lo~largo~de* (PREP), *tendrán~que* (VMfin), *por~el~momento* (ADV), *al~mismo~tiempo* (ADV)
- *cure-dents, gut-ausgearbeitet, Jelzin-Ära, franco-tedesco*
  ×
  padne - li, Tchaj - wan, česko - německý
- *under*, *because*: en:**IN**
- *těch*: cs:**PD** × *tych*: pl:**adj**
- *devátá*: cs:**Cr** × *dziewiąta*: pl:**adj**
- *remotest*: en:**JJS** × **abgelegenste**: de:**ADJA**

InterCorp v. 10

- Preference for electronic sources
- *The Bible*
- Chinese, Japanese
- *treq* – multiword units, query language, English in a ddition to Czech
- Harmonized taxonomy of categories – *Universal Dependencies*?

Universal Dependencies

- A de-facto standard also for morphological categories
- Loss conversion from language-specific tagsets
- Incompatible tokenization
- Alternative: lossless mapping of tagsets onto an ontology of linguistic categories [Chiarcos(2012)]

Linking parallel corpora

- Parallel corpora are useful to speakers of any language
- High synergy in infrastructure and content:
  - Many problems are similar across languages
  - Texts in foreign languages may exist elsewhere
  - Native speakers are the best corpus builders
- Options / Levels of cooperation:
  - Exchange of know-how, tools, texts between centres
  - Virtual integration of content, a common search interface (federated search), a common text dissemination policy
  - A single centre providing coordination and infrastructure for all languages

Grazie mille della vostra attenzione.
Labai dėkoju už dėmesį.
Liels paldies par uzmanību.
Dank u zeer voor uw aandacht.
Dziękuję bardzo Państwu za uwagę.
Muito obrigado pela vossa atenção.
Veľmi pekne vám ďakujem za pozornosť.
Najlepša hvala za vašo pozornost.
Tack så mycket för er uppmärksamhet.
Mange tak for Deres opmærksomhed.
Vielen Dank für Ihre Aufmerksamkeit.
Thank you very much for your attention.
Muchísimas gracias por su atención.
Suur tänu tähelepanu eest.
Oikein paljon kiitoksia mielenkiinnostanne.
Je vous remercie de votre attention.
Nagyon szépen köszönöm a figyelmüket.
Velice vám děkuji za pozornost.

Outline


Belica, C. (2011).
Semantische Nähe als Ähnlichkeit von Kookurenzprofilen.
In A. Abel and R. Zanin, editors, *Korpusinstrumente in Lehre und Forschung*, pages 155–178, Brixen. Bozen-Bolzano University Press.


Chiarcos, C. (2012).
Ontologies of linguistic annotation: Survey and perspectives.
In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 303–310, Istanbul, Turkey. European Language Resources Association (ELRA).


Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014).
The Sketch Engine: ten years on.
Lexicography, **1**(1), 7–36.