
CZECH NATIONAL
CORPUS

CZECH NATIONAL CORPUS

Parallel corpora
in translation and contrastive studies

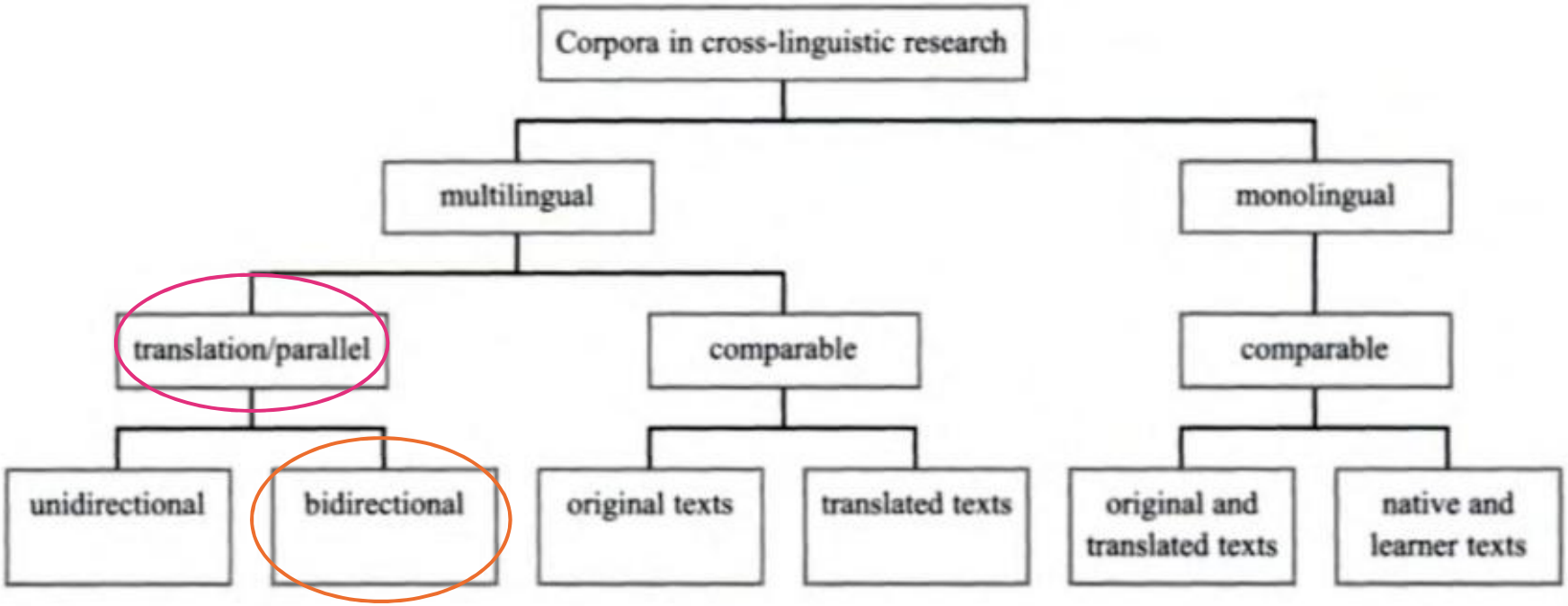
Lucie Chlumská

Faculty of Arts, Charles University in Prague

OUTLINE

1. corpus classification and terminology in TS/CS
2. parallel corpora: objectives and issues
3. InterCorp 9: corpus design
4. languages in contrast based on the parallel corpus

Corpora in TS/CS: terminology



See Granger S., Lerot J. & Petch-Tyson S. (2003) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies. Amsterdam: Rodopi.*

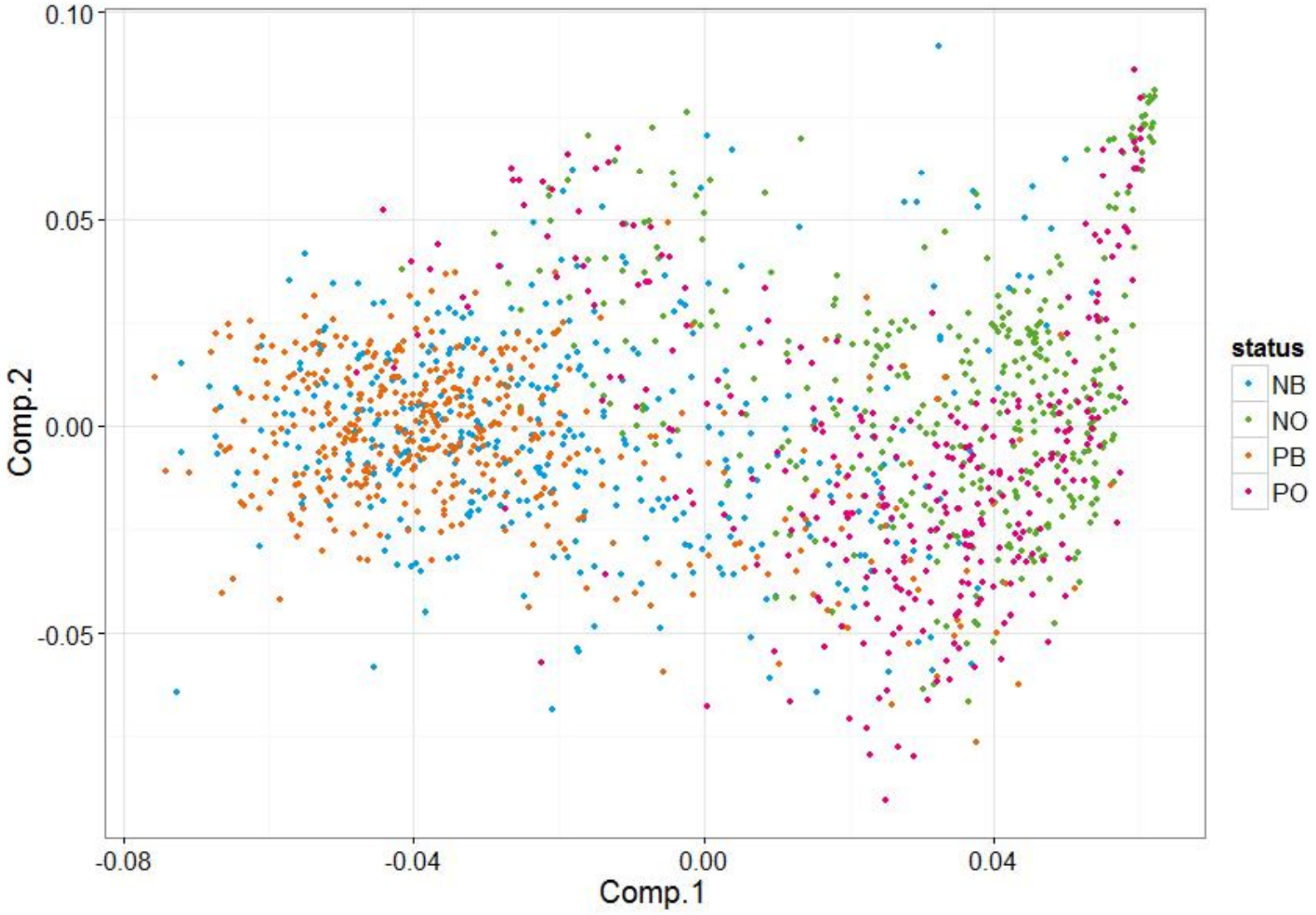
PARALLEL CORPORA

Objectives and issues

- to include originals and their translations
  - segment/sentence alignment, word-to-word allignment?
- to provide a basis for research in TS/CS
- main resource of data for machine translation

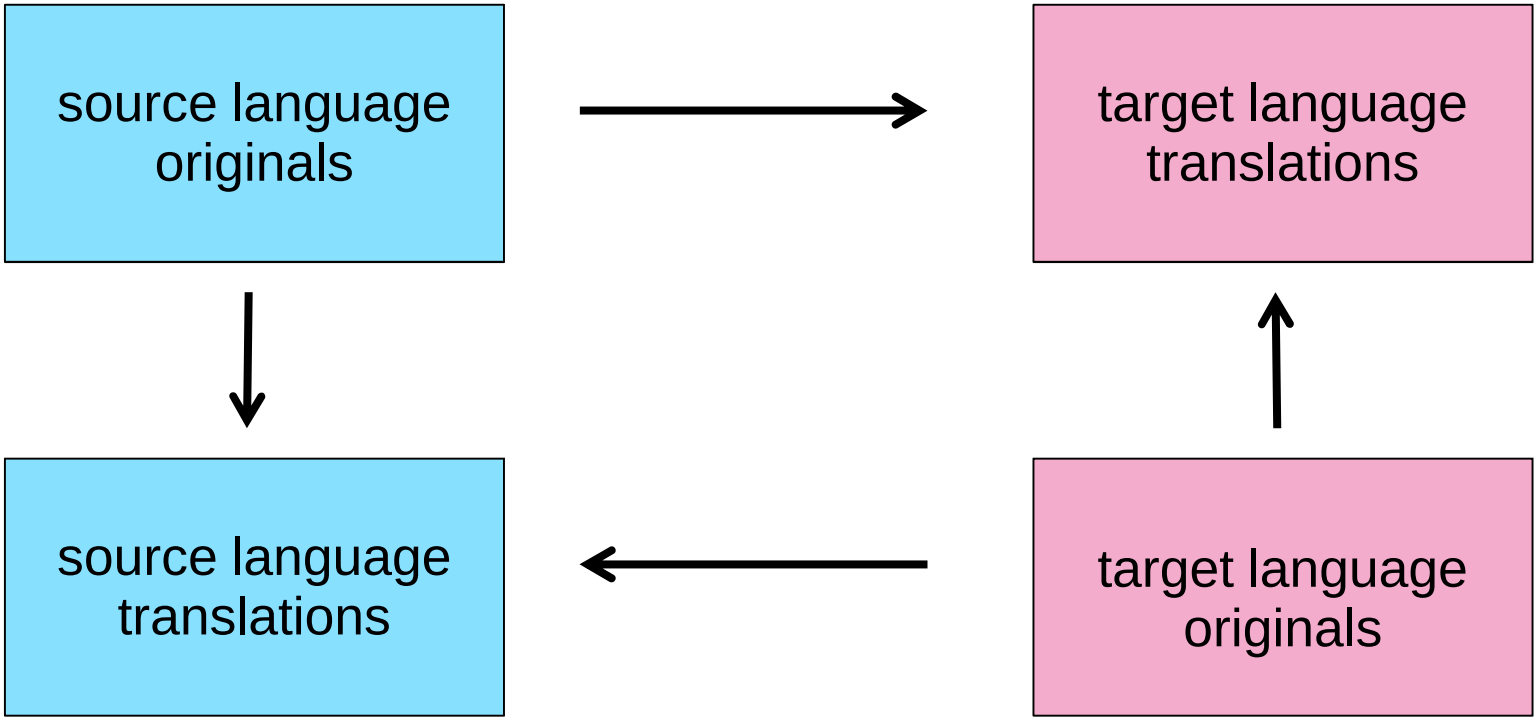
- representativness – genres/text types matter
  - obvious issue in CL: what texts to include? > what translations to include...?
- directionality – small languages vs. big languages
  - different amount of texts translated and available
  - highbrow literature and classics vs. virtually anything is available

PCA: fiction vs. non-fiction



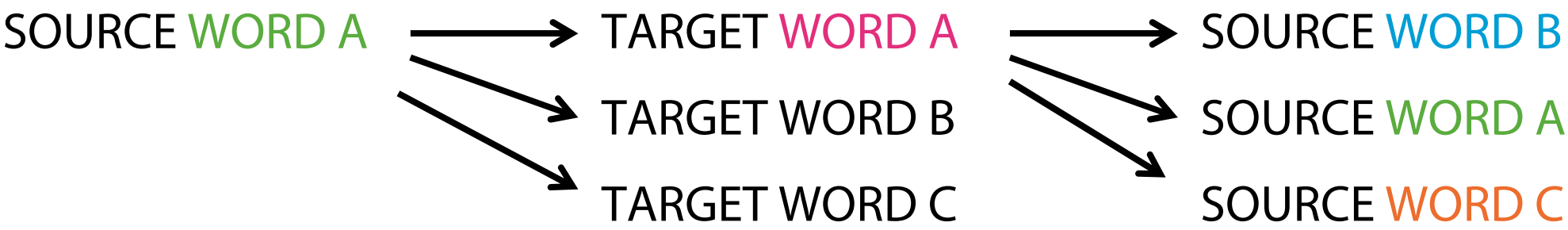
Bidirectional parallel corpus

- same size in both directions > „reciprocal“ (Zanettin 2011)
- both a parallel and comparable corpus (e.g. ENPC) > perfect for the analysis of translation universals (s-universals, t-universals)



Directionality matters

- usually, there is no symmetry in translation equivalence
- ALWAYS DEPENDS ON THE CONTEXT



example:

EN shout > CS křičet > EN scream, shout, yell (EN scream > CS křičet, řvát, ječet)

EN come > CS jít > EN go, come

CS hned > DE gleich > CS stejný, hned, stejně

INTERCORP v.9

Basic information

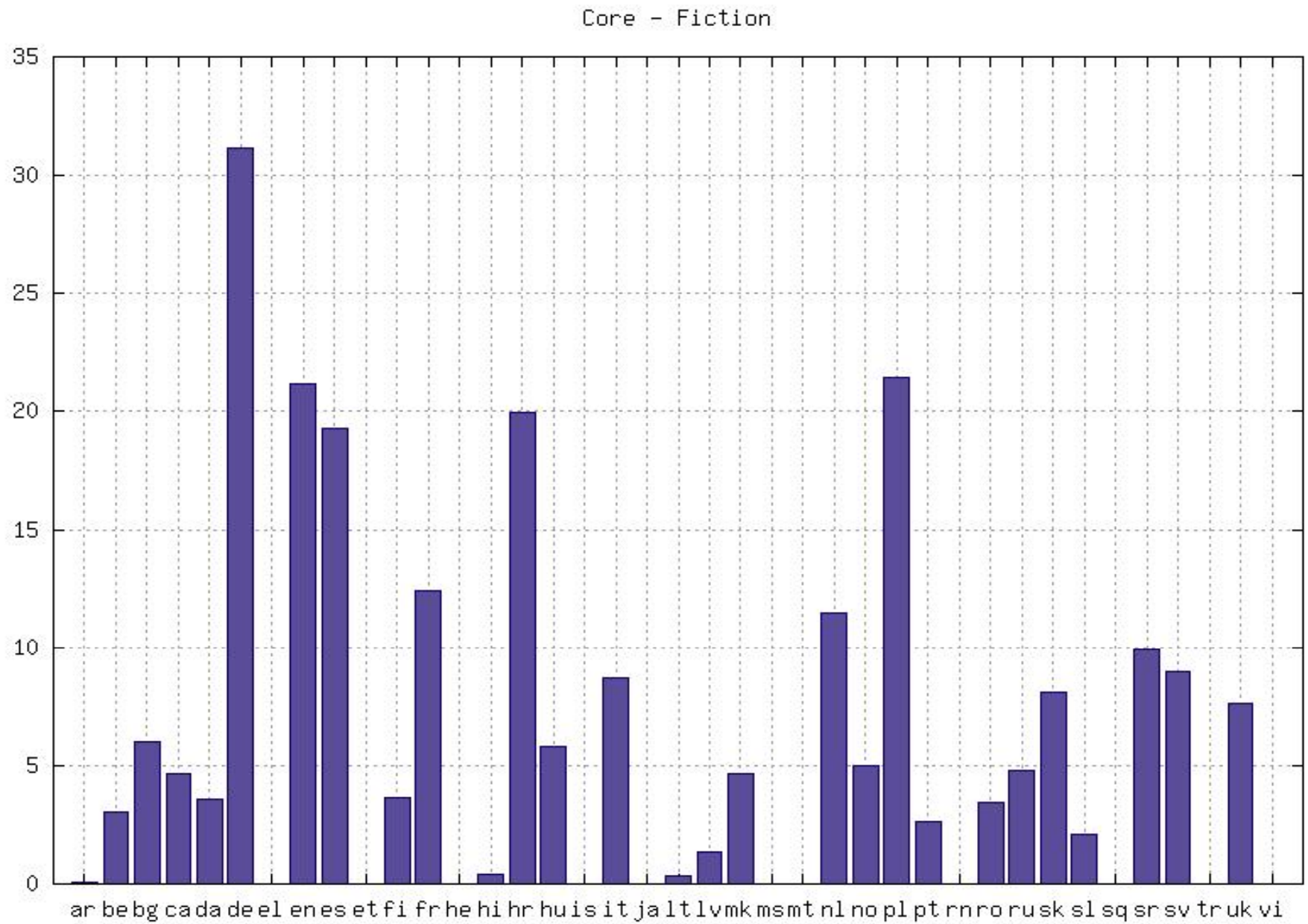
- multilingual parallel corpus focused on Czech (pivot)
- Czech as pivot, sentence/segment alignment
  - word-to-word alignment > used in Treq (treq.korpus.cz)

Name		Czech – core	Czech – collections	other – core	other – collections
Positions	Number of tokens	120,443,181	117,981,673	278,445,878	1,556,840,965
	Number of word forms	96,956,714	89,645,545	231,501,606	1,228,896,294
Structural attributes	Number of documents	1430	5	2,934	89
	Number of div	1,430	111,263	2,934	1,849,184
	Number of sentences	8,308,814	13,588,082	17,210,601	143,478,514
Further information	reference	YES			
	representative	NO			
	publication date	2016			
	foreign languages	39			
	tagged languages	23			
	lemmatized languages	20			

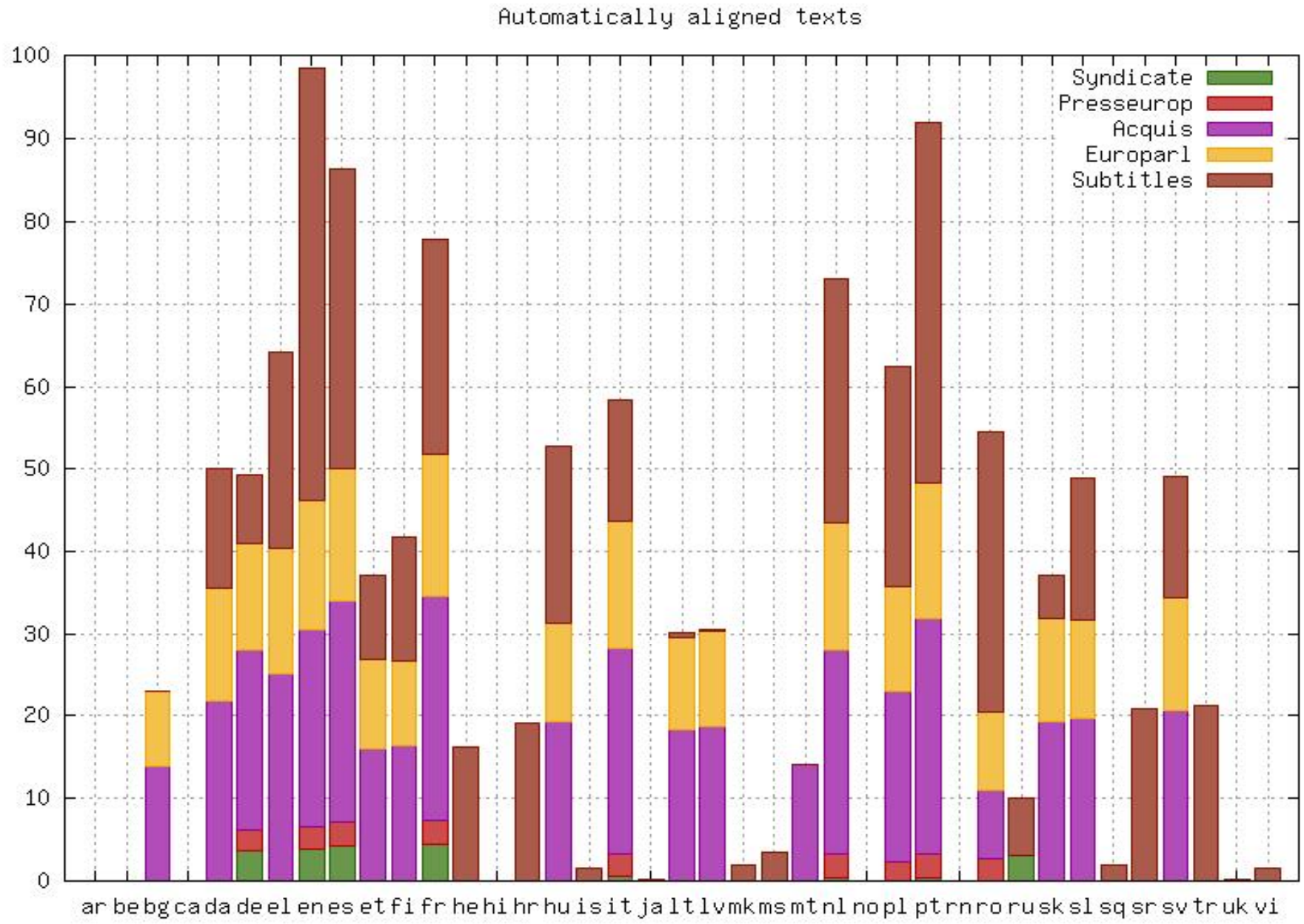
InterCorp 9: design

- currently 39 languages
  - in different proportions, not all are lemmatized and/or tagged
- design: core and collections (incl. subtitles)
  - fiction, manual alignment
  - journalism:
    - Project Syndicate: http://www.project-syndicate.org/
    - PressEurop: http://www.presseurop.eu
  - legal texts in the EU languages:
    - Acquis Communautaire: http://langtech.jrc.ec.europa.eu/JRC-Acquis.html
  - EP (verbatim 2007-2011):
    - Europarl: http://www.statmt.org/europarl/
  - Open Subtitles
    - www.opensubtitles.org

Core


Core - Fiction
35
30
25
20
15
10
5
0
ar be bg ca da de el en es et fi fr he hi hr hu is it ja lt lv mk ms mt nl no pl pt rn ro ru sk sl sq sr sv tr uk vi

Collections



Tags in different languages

Jazyk	Zn.	Lm.	Nástroj	Předl. Det. Adj. Subst.
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

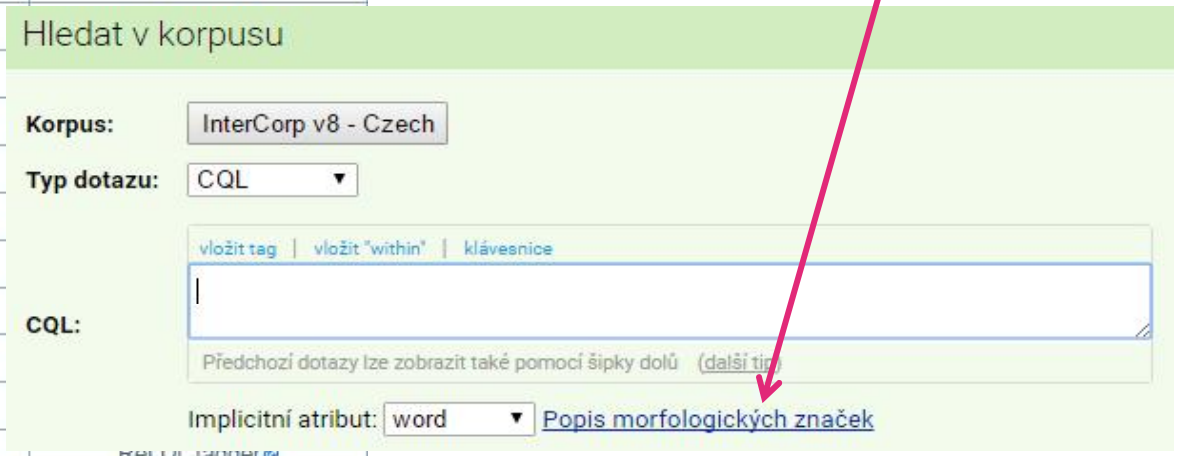
Where to find the tagset description?

Language	Tags	Lemmas	Brief description	Detailed description	Tool
Bulgarian	✓			in English	TreeTagger
Croatian	✓	✓	in English		ReLDI Tagger
Czech	✓	✓	in Czech and in English[2]	in English	Morče
Dutch	✓			in Dutch	TreeTagger
English	✓	✓	in English	in English + additions	TreeTagger
Estonian	✓	✓	in Estonian and English		TreeTagger
Finnish	✓	✓		English[3]	OMorFi+HunPOS
French	✓	✓	in English		TreeTagger
German	✓	✓	in English[4]	in German	RFTagger
Hungarian	✓			in English	HunPos
Icelandic	✓	✓	in English		IceStagger
Italian	✓	✓	in English		
Latvian	✓	✓	in Latvian		
Lithuanian	✓	✓	in Czech and English	in English	
Norwegian	✓	✓	in English and Norwegian		
Polish	✓	✓	in English and Polish	in English	
Portuguese	✓	✓	in Spanish		
Russian	✓	✓	in English	in English[5]	
Slovak	✓	✓	in Slovak	in Slovak	
Slovene	✓	✓	in English and Slovene	in English	
Serbian	✓	✓	in English		ReLDI Tagger
Spanish	✓	✓	in English		TreeTagger
Swedish	✓	✓	in Swedish and English		Stagger

in the Wiki:

http://bit.ly/1bv3ll4

in the KonText interface:



LANGUAGES IN CONTRAST

Examples of use

word-formation

1. EN: *-ridden*, *-laden*

   > meaning? combinations? text types? translations?

2. EN: *Hey , ai n't you that demon-fighting-son-of-a-bitch* ?

   stared up at it with a the-bigger-they-are-the-harder-they-fall expression

   > length? translations?

3. CS: deminutives ending in *–eček*, *-ička*

   > translations? possible equivalents in analytical languages?

Examples of use

grammar

4. EN: present perfect and its counterparts in other languages

   he has never given me a present before vs. he's ~~got(ta)~~, I've ~~been~~ divorced...

   have/has/'s/'ve + any word ~~(been)~~ + past participle ~~(been, got(ta))~~

   > tense? > aspect? > markers?

5. EN: *-ing* clauses – clauses with participle constructions

   Having published a draft of this Regulation, ...

   > transgressives? finite clauses?

6. EN: syntactical feature – disjunct

   Sadly, he came late. *Honestly, I didn't do it.*

Examples of use

pragmatics

7. EN: ...*and stuff*, *sort of*..., *kind of*...
8. CS: *vole* vs. EN: *man*? *dude*? you?

> use? translations? combinations?

lexicon and phraseology

9. proverbs and sayings in different languages

EN: *light as a feather* > in other languages? (ADJ as NOUN)

stylistics / norms of translation

10. verba dicendi

EN: ..., *says Peter/Peter says*. > CS? FI? FR?

Thank you for your attention!

Questions ☺ ?

lucie.chlumska@korpus.cz

Bibliography

- Baker, Mona (1993). Corpus linguistics and translation studies: Implications and applications. In: Baker, M., Francis, G., Tognini-Bonelli, E. (eds.) *Text and Technology: In Honour of John Sinclair.* John Benjamins, Amsterdam-Philadelphia, p. 233-250.
- Corpas, Pastor Gloria & Mitkov, Ruslan & Afzal, Naveed & Pekar, Viktor (2008). Translation universals: Do they exist? A corpus-based NLP study of convergence and simplification. *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA-08).*
- Laviosa-Braithwaite, Sara (1996). Investigating Simplification in English Comparable Corpus of Newspaper Articles. Daniel Berzsenyi College Printing Press Szombathely.
- Laviosa, Sara (1998 Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta: Translator's Journal.* Vol. 43, No. 4, p. 557-571.
- Mihăilă, Claudiu (2010). Translation Studies: Simplification and Explicitation Universals. Available at: http://www.slideshare.net/claudiumihaila/uaic-3801394.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

NON-TYPICAL PATTERNS AND COLLOCATIONS

N-grams: extraction

- **3**-grams & **4**-grams (strings of 3-4 words, excl. punctuation)

  1. automatically generated list of n-grams from Jerome
  2. comparison of relative freqencies in T and N
  3. selection of the most different ones (occurring in one of the subcorpus only, outliers etc.)
  4. manual sorting out of irrelevant results (personal names, text-related phrases etc.)

N-gramy: typical in translations

- 3-grams:
  - *Co to sakra, Děláš si legraci, to tak líto, je mi líto, mi to líto, ani v nejmenším, Zkrátka a dobře...*
- 4-grams:
  - *o čem to sakra, To je v pořádku, je to v pořádku, že je v pořádku, Všechno bude v pořádku, Moc mě to mrzí, až do morku kostí, co do činění s, Pokud jde o mě, Podle mě je to, Pro všechno na světě...*

interference from EN (*v pořádku, líto, mrzí...*)

N-gramy: typical in non-translations

- 3-grams:
  - *jen a jen, další a další, v neposlední řadě, v té době...*
- 4-grams:
  - *stále nové a nové, čím dál tím méně, čím dál tím více, mezi nebem a zemí, a tak není divu, jako jeden z mála, od rána do noci...*

repetitions, different phrasemes...