Stylometry in Literary Translation via Universal Dependencies: Finally Breaking the Language Barrier?

Silvie Cinková, Jan Rybicki

Stylometry in Literary Translation

- Stylometry: detect author's personal signal in text
- most frequent words, most frequent n-grams
  - function words (supposedly beyond author's control)
- cluster analysis, network analysis of authors
- All documents must be in the same language!

authors	books per author	translations per book/author	signal strength
n	1	n	translators_selfauthors > translators pure
n	n	n	authors > translators
n	n	2	translators > authors

stylo...

- requires a collection of plain texts
- for each document: frequencies of unique strings (words? lemmas? something else?)

1	Emma	NOUN	Case=Nom|Gender=Fem|Number=Sing	20	nsubj
2	Woodhouse	NOUN	Case=Gen|Gender=Fem|Number=Sing	1	nmod
3	,	PUNCT	NA	4	punct
4	osóbka	NOUN	Case=Nom|Gender=Fem|Number=Sing	1	conj
5	przystojna	ADJ	Case=Nom|Degree=Pos|Gender=Fem|Number=Sing	4	amod

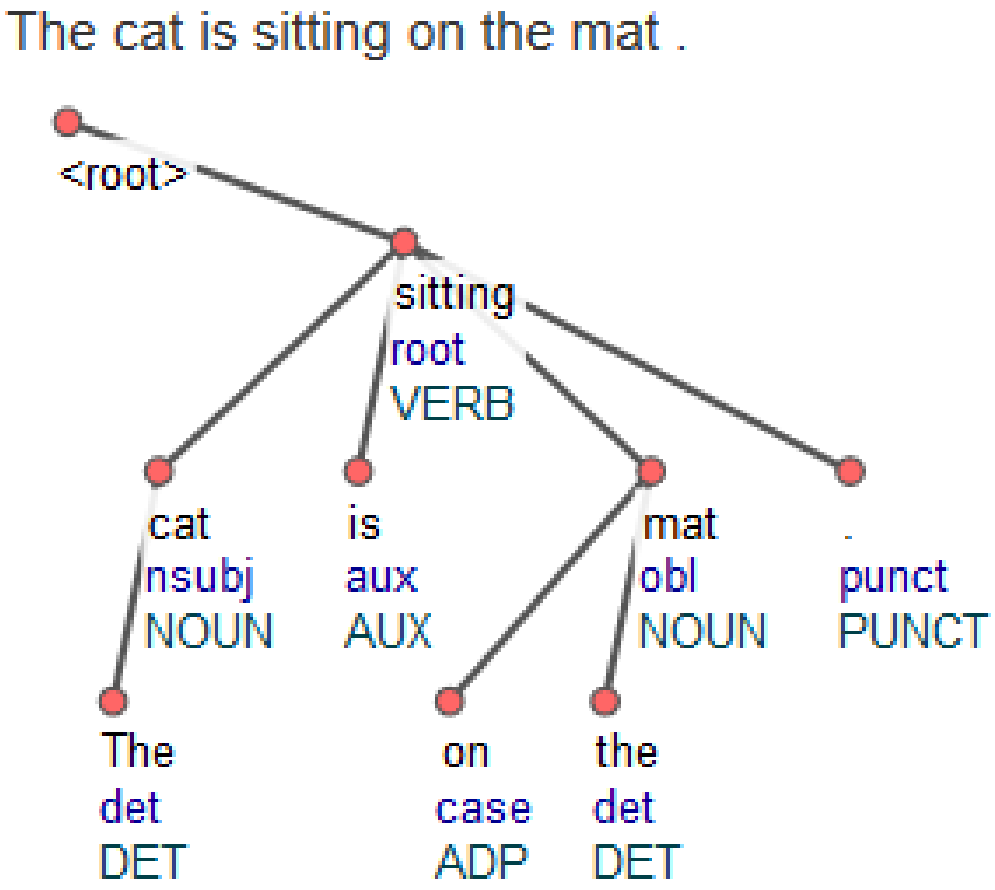
universaldependencies.org

Universal POS (parts of speech)		
ADJ	adjective	*good*
ADP	adposition	*before, ago, in*
ADV	adverb	*fortunately*
AUX	auxiliary verb	*be, have*
CCONJ	coord. conjunction	*and, but*
DET	determiner	*this, a, both*
INTJ	interjection	*oh, thanks*
NOUN	noun, cf. PROPN	*woman*
NUM	numeral	*two, 12, VII*
PRON	pronoun	*you*
PART	particle	*indeed, just*
PUNCT	punctuation	*,;!.*
SCONJ	subord. conjunction	*because*
SYM	symbol	*$,%*
VERB	verb	*speak*

Feature	Values
Number	Plur, Sing, ...
Animacy	Anim, Inan
Gender	Masc, Fem, ...
VerbForm	Inf, Fin, Part
Definite	Def, Indef
Voice	Act, Pass,...

form	lemma	upos	features
Her	she	PRON	Gender=Fem|Number=Sing|Person=3|Poss=Yes|PronType=Prs
diamonds	diamond	VERB	Number=Pl
blazed	blaze	VERB	Tense=Past|VerbForm=Part
out	out	ADP	_

Syntactic Dependencies



Data Preparation

- all texts parsed with the UDPipe parser

id	form	lemma	upos	xpos	features	head ID	synt. depen-dency
# text = The quick brown fox jumps over the lazy dog.							
1	The	the	DET	DT	Definite=Def|PronType=Art	4	det
2	quick	quick	ADJ	JJ	Degree=Pos	4	amod
3	brown	brown	ADJ	JJ	Degree=Pos	4	amod
4	fox	fox	NOUN	NN	Number=Sing	5	nsubj
5	jumps	jump	VERB	VBZ	Mood=Ind|Number=Sing|Person=3|Tense=Pres		
6	over	over	ADP	IN	_	9	case
7	the	the	DET	DT	Definite=Def|PronType=Art	9	det
8	lazy	lazy	ADJ	JJ	Degree=Pos	9	amod
9	dog	dog	NOUN	NN	Number=Sing	5	nmod
10	.	.	PUNCT	.	_	5	punct

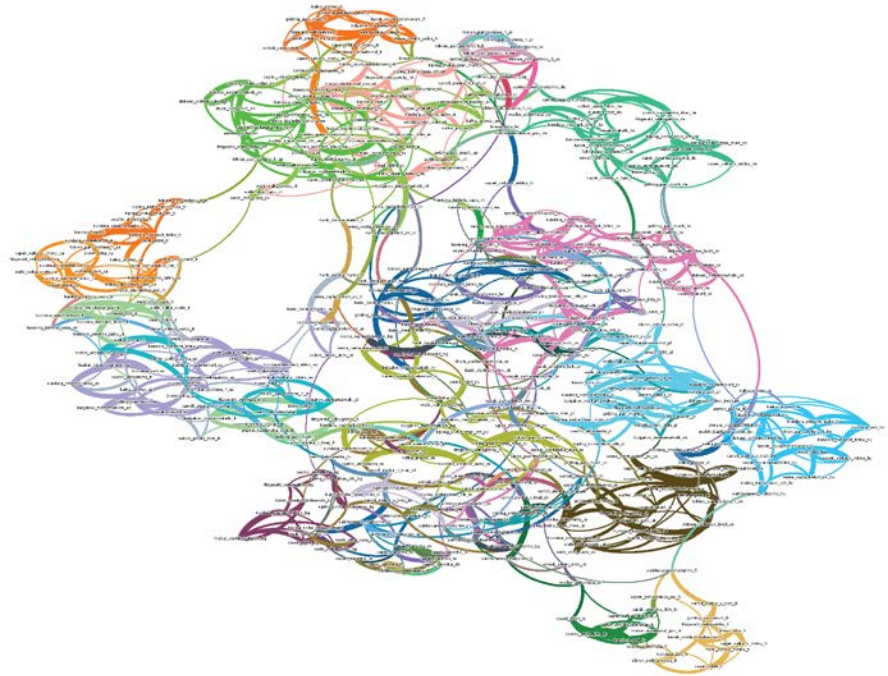
Data Preparation

- some columns stripped

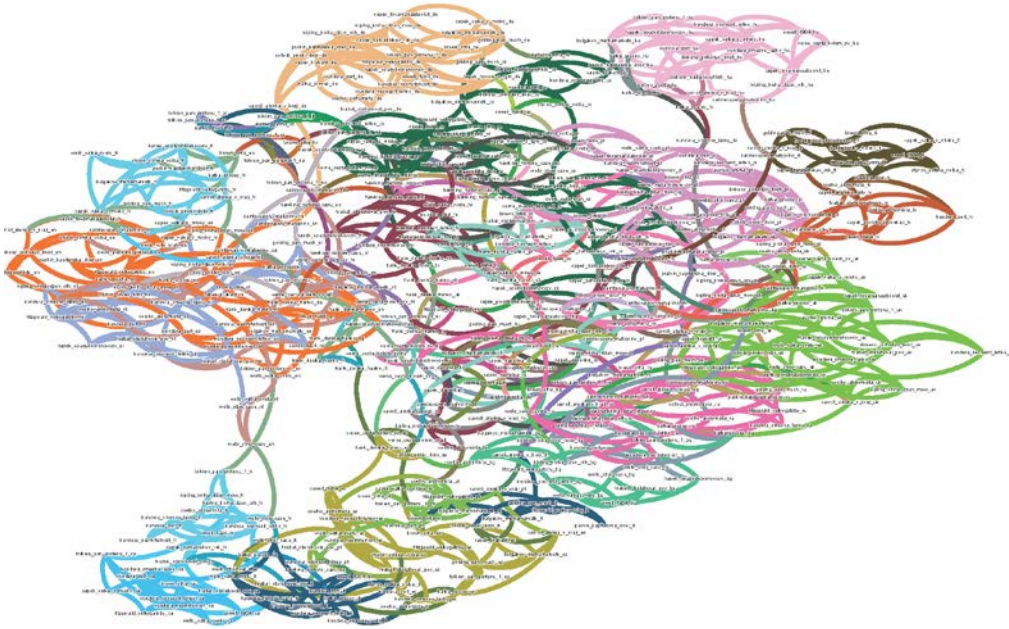
id	form	lemma	upos	xpos	features	head ID	synt. depen-dency
# text = The quick brown fox jumps over the lazy dog.							
1	The	the	DET	DT	Definite=Def|PronType=Art	4	det
2	quick	quick	ADJ	JJ	Degree=Pos	4	amod
3	brown	brown	ADJ	JJ	Degree=Pos	4	amod
4	fox	fox	NOUN	NN	Number=Sing	5	nsubj
5	jumps	jump	VERB	VBZ	Mood=Ind|Number=Sing|Person=3|Tense=Pre		
6	over	over	ADP	IN	_	9	case
7	the	the	DET	DT	Definite=Def|PronType=Art	9	det
8	lazy	lazy	ADJ	JJ	Degree=Pos	9	amod
9	dog	dog	NOUN	NN	Number=Sing	5	nmod
10	.	.	PUNCT	.	_	5	punct

Stylometric Analysis - POS

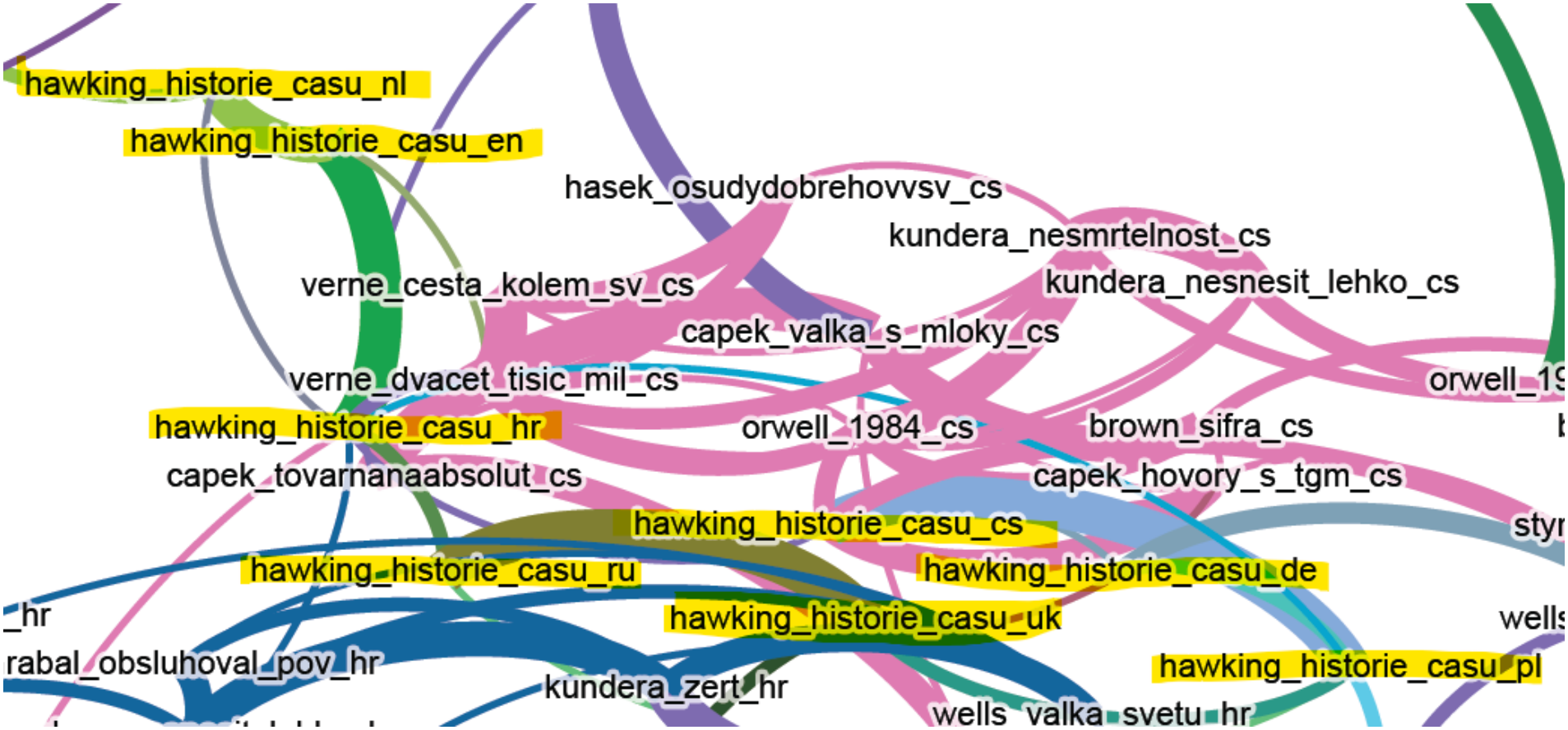
UPOS



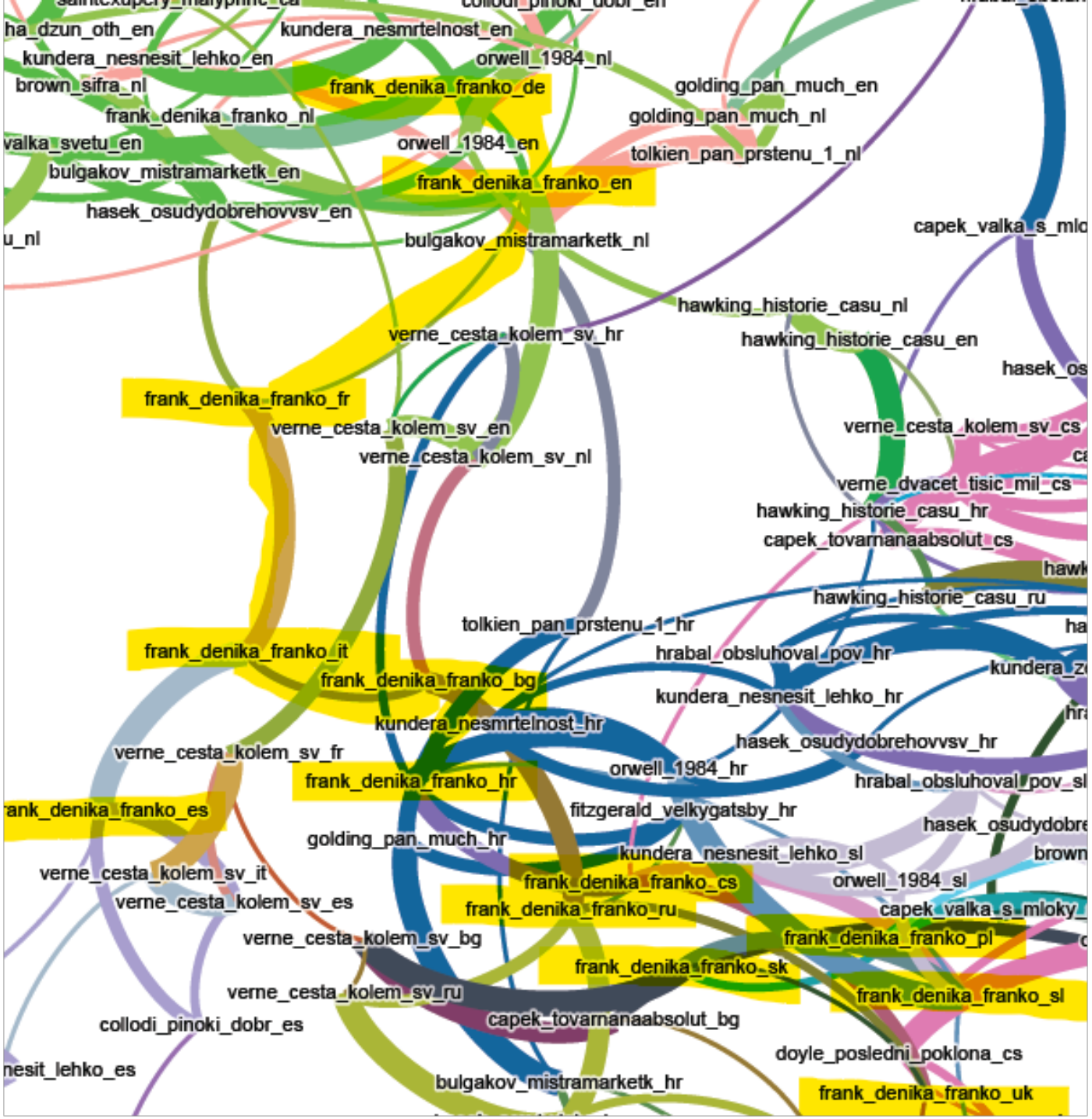
FULL TAGS



Full Tags - Stephen Hawking



Full Tags - Anne Frank



Full Tags - Karel Čapek: Dášeňka

Source Data

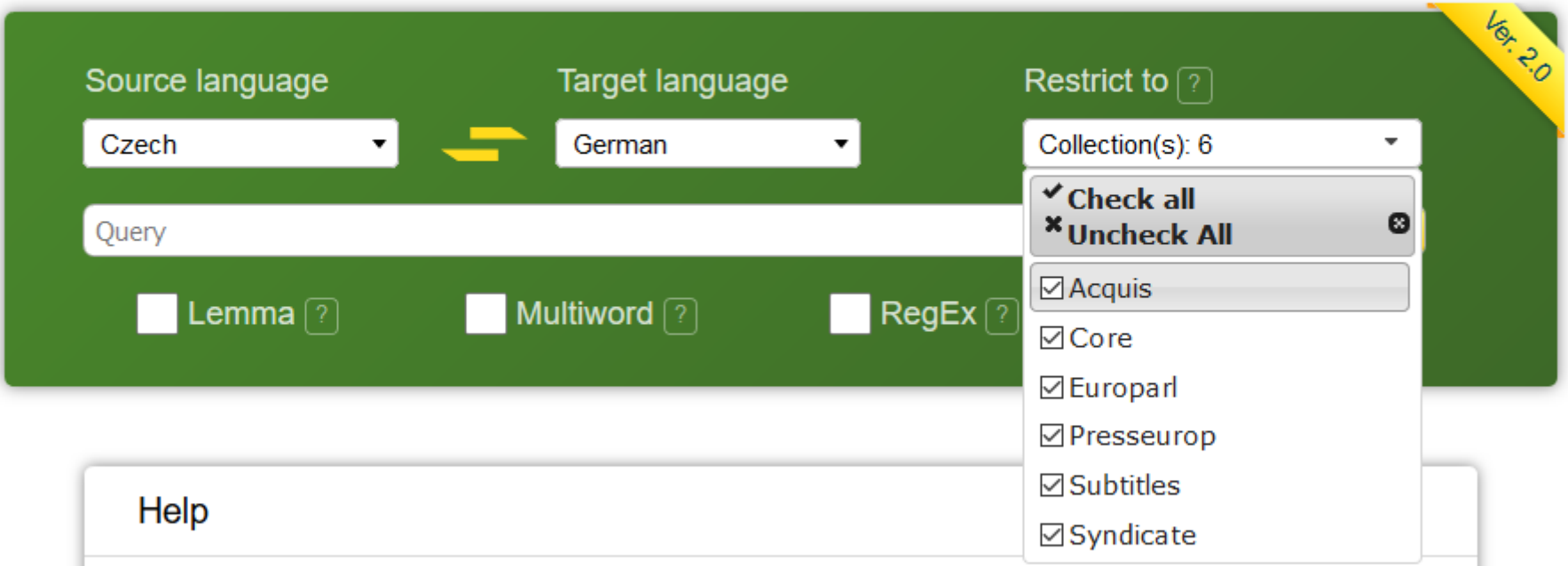
27 (54) CS and DE texts (Czech National Corpus - InterCorp)

- Brown-Da Vinci Code
- Bulgakov-Master and Margarethe
- Capek-Dasenka
- ***Capek-Interviews with TGM***
- Capek-Krakatit
- Capek-The Absolute at Large
- Capek-Salamander War
- *Capek-Gardener's Year*
- Carroll-Alice in Wonderland
- Coelho-Alchymist
- Collodi-Pinocchio's Adventures
- Fitzgerald-The Great Gatsby
- Frank-Anne Frank's diary
- Golding-Lord of Flies
- Hasek-Good Soldier Schwejk
- ***Hawking-History of Time***
- Hrabal-I Served the King of England (tales)
- Kafka-The Process
- Kipling-The Jungle Book - Mowgli
- Kipling-The Jungle Book - Other tales
- Kundera-Immortality
- Kundera-Unbearable Lightness of Being
- Kundera-The Joke
- Orwell-1984
- Pushkin-Captain's Daughter
- SaintExupery-The Little Prince
- Tolkien-Lord of the Rings 1

Adding Pseudolemmas from a Czech-German glossary

Treq
TRANSLATION
EQUIVALENTS
DATABASE

CZECH NATIONAL CORPUS



Integrated Pseudolemmas

glossary

freq	CS	DE	flemma
182686	"	«	L00009
163642	?	?	L00010
139614	být	sie	L00011
139120	ten	die	L00012
128976	že	dass	L00013
124208	který	die	L00014
118355	já	sie	L00015
112098	na	auf	L00016
107206	s	mit	L00017
91365	!	!	L00018

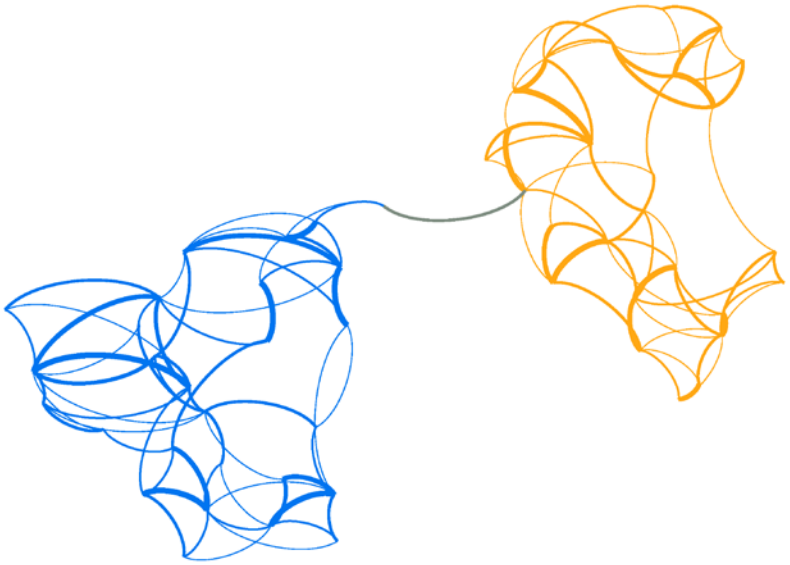
pseudolemmas in German text

sent id	flemma	upos	features	deprel
doc4	L00044	DET	Case=Acc Number=Sing	det
doc4	L00263	NOUN	Case=Acc Number=Sing	obj

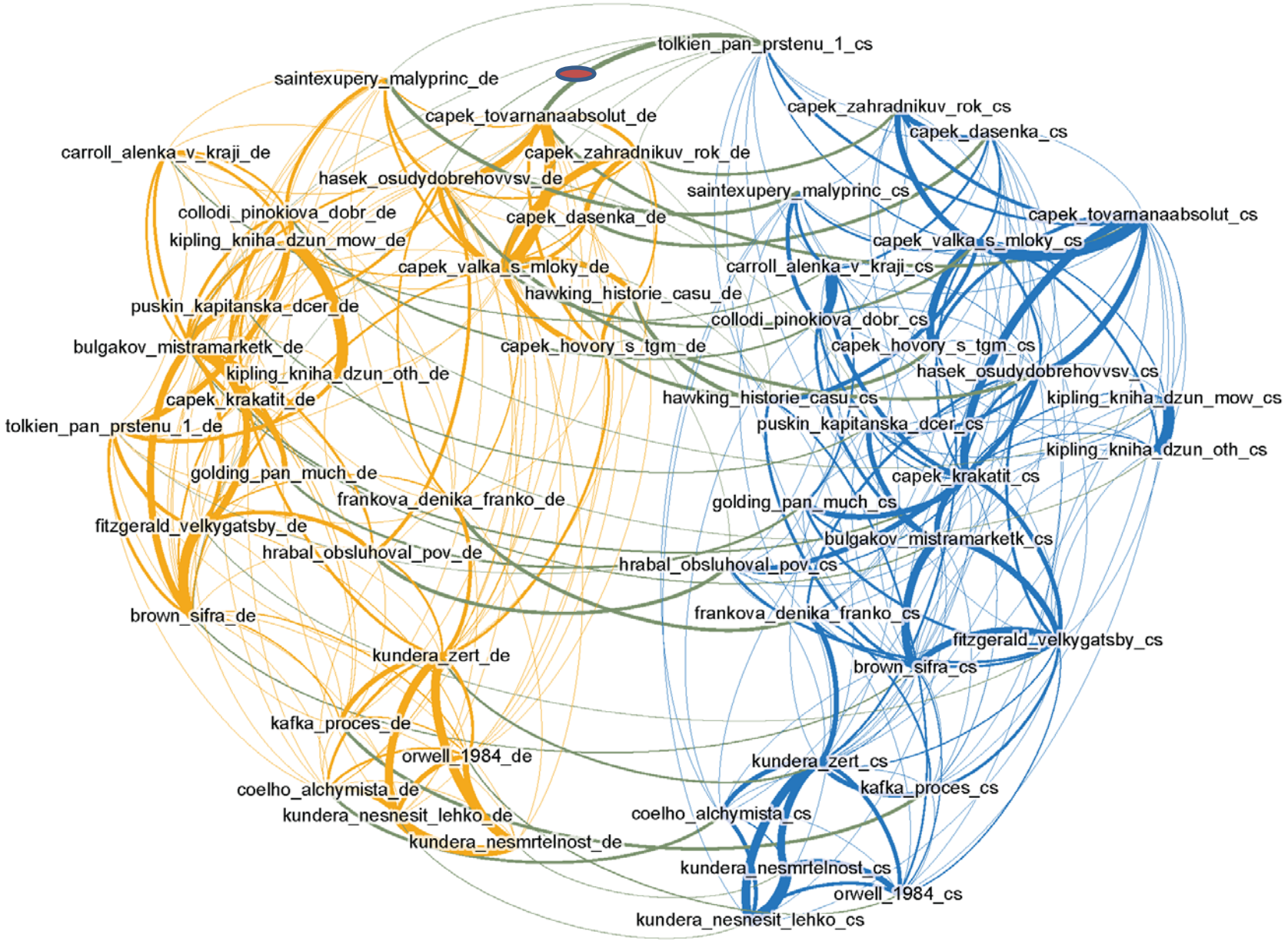
automatic alignment in CNC InterCorp: VERY NOISY!

Effect of Pseudolemmas

Full tags without pseudolemmas



pseudolemmas + POS


tolkien_pan_prstenu_1_cs
saintexupery_malyprinc_de
capek_tovarnanaabsolut_de
capek_zahradnikuv_rok_cs
capek_dasenka_cs
carroll_alenka_v_kraji_de
capek_zahradnikuv_rok_de
hasek_osudydobrehovvsv_de
saintexupery_malyprinc_cs
collodi_pinokiova_dobr_de
capek_dasenka_de
capek_tovarnanaabsolut_cs
kipling_kniha_dzun_mow_de
capek_valka_s_mloky_cs
capek_valka_s_mloky_de
carroll_alenka_v_kraji_cs
hawking_historie_casu_de
puskin_kapitanska_dcer_de
collodi_pinokiova_dobr_cs
bulgakov_mistramarketk_de
capek_hovory_s_tgm_de
capek_hovory_s_tgm_cs
kipling_kniha_dzun_oth_de
hasek_osudydobrehovvsv_cs
capek_krakatit_de
hawking_historie_casu_cs
kipling_kniha_dzun_mow_cs
tolkien_pan_prstenu_1_de
puskin_kapitanska_dcer_cs
kipling_kniha_dzun_oth_cs
golding_pan_much_de
capek_krakatit_cs
frankova_denika_franko_de
golding_pan_much_cs
fitzgerald_velkygatsby_de
bulgakov_mistramarketk_cs
hrabal_obsluhoval_pov_de
hrabal_obsluhoval_pov_cs
brown_sifra_de
frankova_denika_franko_cs
fitzgerald_velkygatsby_cs
kundera_zert_de
brown_sifra_cs
kafka_proces_de
orwell_1984_de
kundera_zert_cs
coelho_alchymista_de
kafka_proces_cs
kundera_nesnesit_lehko_de
coelho_alchymista_cs
kundera_nesmrtelnost_de
kundera_nesmrtelnost_cs
orwell_1984_cs
kundera_nesnesit_lehko_cs

Results

Tagging	Attribution success
POS	3.7%
POS + features	3.7%
pseudolemmas	3.7%
pseudolemmas + POS + syntactic dependencies	10.2%
pseudolemmas + POS + features + syntactic dependencies	16.7%
POS + features + syntactic dependencies	**20.3%**
pseudolemmas + POS + features	56.7%
pseudolemmas + POS	**95.6%**

Future work

- **try multilingual machine translation on bigger data!**
- baseline with full forms (totally language-specific): comparison of clustering of authors within languages with other setups
- check out different language pairs with pseudolemmas