Applying an automatic FTD classifier to the annotation of the GICR corpus

M Bulygin, S Sharoff

Centre for Translation Studies
University of Leeds

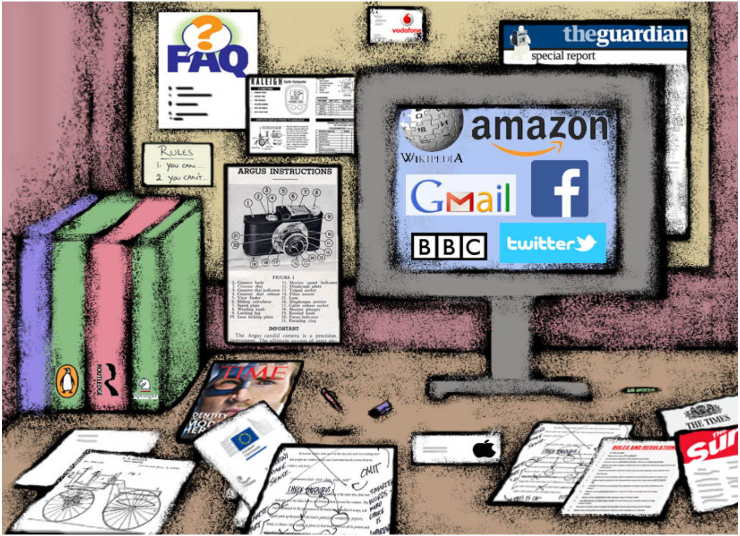
<2019-05-30>

UNIVERSITY OF LEEDS

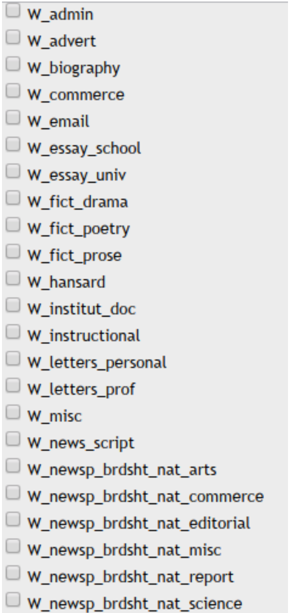
Outline

1. Genres of everyday life
   - Traditional corpora vs Web
2. Functional Text Dimensions
   - Typologies of genre labels
   - Reliable topology of genres
3. Automatic genre annotation
   - Machine learning approaches
   - Results on test set
   - Application of genres to GICR
   - Comparison to neural methods
4. Conclusions

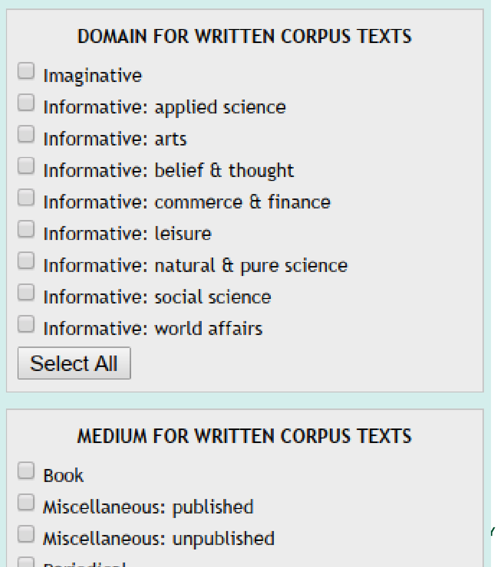
Genres of everyday life



Querying traditional corpora: BNC



- □ W_admin
- □ W_advert
- □ W_biography
- □ W_commerce
- □ W_email
- □ W_essay_school
- □ W_essay_univ
- □ W_fict_drama
- □ W_fict_poetry
- □ W_fict_prose
- □ W_hansard
- □ W_institut_doc
- □ W_instructional
- □ W_letters_personal
- □ W_letters_prof
- □ W_misc
- □ W_news_script
- □ W_newsp_brdsht_nat_arts
- □ W_newsp_brdsht_nat_commerce
- □ W_newsp_brdsht_nat_editorial
- □ W_newsp_brdsht_nat_misc
- □ W_newsp_brdsht_nat_report
- □ W_newsp_brdsht_nat_science



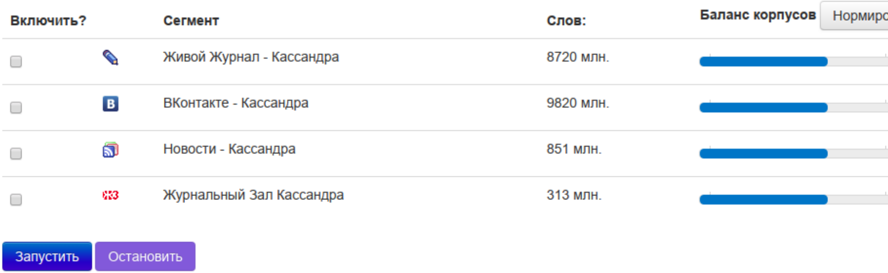
DOMAIN FOR WRITTEN CORPUS TEXTS

- □ Imaginative
- □ Informative: applied science
- □ Informative: arts
- □ Informative: belief & thought
- □ Informative: commerce & finance
- □ Informative: leisure
- □ Informative: natural & pure science
- □ Informative: social science
- □ Informative: world affairs

MEDIUM FOR WRITTEN CORPUS TEXTS

- □ Book
- □ Miscellaneous: published
- □ Miscellaneous: unpublished
- □ Periodical

Querying Web corpora: GICR



Включить?		Сегмент	Слов:	Баланс корпусов Нормиро
☐		Живой Журнал - Кассандра	8720 млн.	
☐	B	ВКонтакте - Кассандра	9820 млн.	
☐		Новости - Кассандра	851 млн.	
☐	ЖЗ	Журнальный Зал Кассандра	313 млн.	

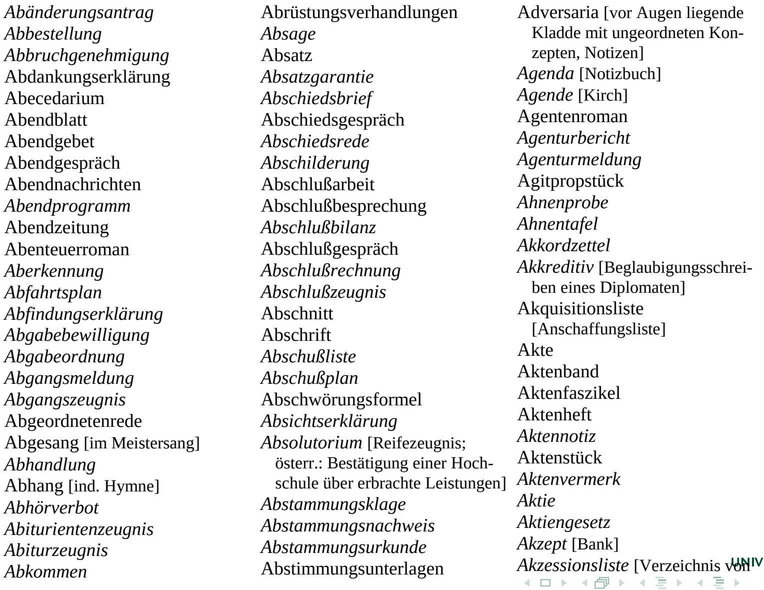
Typologies of genre labels

Brown, LOB, LCMC 500 samples, 2000 words in each, belonging to 15 genres: A) Press: reportage, B) Press: editorial, C) Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) SciFi, N) Adventure ...

BNC about 4,000 texts with classification into 70 genres (*ac.med*, *ac.tech*, *nonac.tech*, *ac.socsci*, *ac.politlaw*, *news.brd.nat.art*, *news.brd.nat.com*...), medium (*book*, *periodical*, *ephemeral*...), audience...

BL for fiction Adventure stories, Detective stories, Picaresque stories, Robinsonades, Sea stories, Spy stories, Thrillers, Allegories, Didactic fiction, Fables, Parables, Alternative histories, Dystopias, Bildungsromane, Arthurian romances ...

6,500 genres from (Adamzik, 1995)



Abänderungsantrag
Abbestellung
Abbruchgenehmigung
Abdankungserklärung
Abecedarium
Abendblatt
Abendgebet
Abendgespräch
Abendnachrichten
Abendprogramm
Abendzeitung
Abenteuerroman
Aberkennung
Abfahrtsplan
Abfindungserklärung
Abgabebewilligung
Abgabeordnung
Abgangsmeldung
Abgangszeugnis
Abgeordnetenrede
Abgesang [im Meistersang]
Abhandlung
Abhang [ind. Hymne]
Abhörverbot
Abiturientenzeugnis
Abiturzeugnis
Abkommen

Abrüstungsverhandlungen
Absage
Absatz
Absatzgarantie
Abschiedsbrief
Abschiedsgespräch
Abschiedsrede
Abschilderung
Abschlußarbeit
Abschlußbesprechung
Abschlußbilanz
Abschlußgespräch
Abschlußrechnung
Abschlußzeugnis
Abschnitt
Abschrift
Abschußliste
Abschußplan
Abschwörungsformel
Absichtserklärung
Absolutorium [Reifezeugnis; österr.: Bestätigung einer Hochschule über erbrachte Leistungen]
Abstammungsklage
Abstammungsnachweis
Abstammungsurkunde
Abstimmungsunterlagen

Adversaria [vor Augen liegende Kladde mit ungeordneten Konzepten, Notizen]
Agenda [Notizbuch]
Agende [Kirch]
Agentenroman
Agenturbericht
Agenturmeldung
Agitpropstück
Ahnenprobe
Ahnentafel
Akkordzettel
Akkreditiv [Beglaubigungsschreiben eines Diplomaten]
Akquisitionsliste [Anschaffungsliste]
Akte
Aktenband
Aktenfaszikel
Aktenheft
Aktennotiz
Aktenstück
Aktenvermerk
Aktie
Aktiengesetz
Akzept [Bank]
Akzessionsliste [Verzeichnis von

Functional Text Dimensions (Sharoff, 2018)

A08 news To what extent does the text provide an informative report of recent events? (newswires)

A01 argum To what extent does the text try to persuade the reader? (argumentative blog entries or newspaper opinion columns)

A17 review To what extent does the text evaluate a specific entity? (product reviews)

A11 personal To what extent does the text report from a first-person point of view? (diary-like blog entries)

Rating Levels

0	none;	Ignore hesitations
0	slightly;	⇑
.5	somewhat or partly;	⇓
1	strongly.	Emphasise confident judgements

Full list of principal FTDs

Code	Abbrev	Prototypes
A1	argum	Argumentative blogs or opinion pieces
A4	fictive	Fiction, myths, film plots
A7	instruct	Tutorials or FAQs
A8	news	Reporting newswires
A9	legal	Laws, contracts, copyrights
A11	personal	Diary-like blog entries
A12	promo	Adverts
A13	prop	Propaganda
A14	academ	Academic research papers
A16	info	Encyclopedic articles or text books
A17	eval	Reviews of products, locations or performance
A20	appell	Small ads, requests

Training corpus

Data set	Documents	Words
Training data	1800	2249818
Testing data	140	163923
Total	1940	2413741

FTD	A1	A4	A7	A8	A9	A11
Size	0.14	0.05	0.04	0.25	0.05	0.12

FTD	A12	A14	A16	A17
Size	0.17	0.1	0.12	0.09

Features for genre annotation

- Word features: topics or genres?
- Mixing words and POS (Baroni and Bernardini, 2006)
  {*It won the SCBWI Golden Kite Award for best nonfiction book of 1999 and has sold about 50,000 copies.* }
  {`it won the PROPN ADJ NOUN NOUN for best NOUN NOUN of [#] and has sold about [#] NOUN.` }
- Character n-grams generalise well (Sharoff et al., 2010)
  {day_ =*yesterday,Monday,Tuesday,Saturday,Sunday*; *d_ by* Passive; *sly_* Adverbs}
- N-gram size and capitalisation:
  2-5 grams: наро → народ, международный, народной, народности, народно-освободительной, коммунаров, всенародно
  lowercasing: Только коммунисты могут вернуть власть НАРОДУ!

Algorithms

- Blackbox of Neural Networks
- SVM in the Python scikit-learn library
- Choice of kernels:
  - rbf (radial basis function)
  - poly (polynomial),
  - sigmoid,
  - linear (← our choice)
- Standard classification instead of topology

Accuracy for categories

Model		A1	A4	A7	A8	A9	A11	A12	A14	A16	A17
5gr-C1	P	0.83	1	1	0.97	1	0.83	1	0.92	0.78	1
	R	0.45	0.67	0.5	0.86	0.86	0.31	0.88	0.92	0.41	0.56
3gr-C10	P	0.79	0.75	1	0.94	0.86	0.38	0.89	0.93	0.53	0.92
	R	0.58	1	0.5	0.92	0.86	0.31	0.85	0.93	0.47	0.65
5gr-C10	P	0.91	0.6	1	0.98	0.86	0.46	1	1	0.8	0.92
	R	0.77	1	0.5	0.9	0.86	0.38	0.9	0.93	0.42	0.71
3gr-C100	P	0.75	0.75	1	0.94	0.86	0.33	0.89	0.93	0.56	0.92
	R	0.58	1	0.5	0.92	0.86	0.31	0.85	0.93	0.47	0.65
5gr-C100	P	0.86	0.6	1	0.98	0.86	0.46	1	1	0.8	0.92
	R	0.73	1	0.5	0.9	0.86	0.38	0.9	0.93	0.42	0.71
3gr-C10-fs	P	0.67	0.75	1	0.93	1	0.5	0.87	1	0.53	0.92
	R	0.55	1	0.5	0.86	0.86	0.31	0.81	1	0.47	0.69
5gr-C10-fs	P	0.68	0.5	1	0.95	1	0.36	0.88	0.86	0.53	0.92
	R	0.59	1	0.5	0.88	0.86	0.25	0.88	1	0.53	0.69

N-gram features per categories

FTD	Top features	Words
A1	соци, оказа, прич, ителя, наро	социальной, оказания, причем, представ международного, богатства, показатели, доказател оказались, причин, заместителя, народа
A4	его , казал, глаз, , не	его, сказал, казалось, показал, некоторые
A7	нстру, добав, форма, вас , если , если, запр	конструкции, добавить, год, информации, желание, если, запрос, инструментов, порой, запрещено
A8	ября , новы, сказа, сообщ, моск, явил	сентября, новых, сказал, московских, сообщения, з октября, основы, появились
A9	ветст, должн, мать , стать, федер	должны, соответствии, статьи, заказ, принимать, р ответственности, должностных, федерального,
A11	много, лет , свое, нас , перед, лись , ‘ меня	оказались, появились, остались, находились, изв совместно
A12	азмер , крас, овый , для , прод, наком, ство	красоты, красный, красивый, новый, продукции, количество, большинство
A14	она , зада, ений , больш,) . , расс, ости ,	задачи, рассмотрения, отношений, решений, изме деятельности
A16	начал, изма , нный , . в , прин, века , жела	начала, механизма, данный, принять, человека, же организма, современный единственный
A17	хотя , смотр, разу , овски, мало, без , стати, книг	рассмотрения, сразу, московский, кстати, разумеется, банковские

Application of genres to GICR

Segment	A1	A4	A7	A8	A9	A11	A12	A14	A16	A17
LiveJo	39.0	2.0	0.3	9.0	0.2	42.0	1.0	0.1	2.0	5.0
mail.ru	52.0	2.0	1.0	0.4	0.0	39.0	1.0	0.0	0.6	4.0
russ.ru	16.0	33.0	0.0	0.3	1.0	24.0	0.0	6.0	14.0	5.0
News	4.0	0.0	0.0	92.0	0.3	0.3	0.2	0.2	2.0	0.4
Vk.com	71.0	4.0	0.3	1.0	0.1	19.0	3.0	0.0	1.0	0.7
GICR	45.0	2.0	0.7	11.0	0.2	30.0	3.0	0.5	2.0	5.0

Interpretation of results

- A1 (argumentative) vs A8 (newswires)
- MultiDimensional Analysis (Biber, 1988)
  Past tense verbs, abstract nouns, hedging

Features	A1	A8
Present tense:	0.03490829	0.02912898
Past tense:	0.01967835	0.03572108

Comparison to neural methods (Kunilovskaya, Sharoff, fc)

- mixed word and POS, BiLSTM, Attention

Model		A1	A4	A7	A8	A9	A11	A12	A
5gr-C10-fs	P	0.68	0.5	1	0.95	1	0.36	0.88	0.8
	R	0.59	1	0.5	0.88	0.86	0.25	0.88	
BiLSTM,Attn	AP	0.70	0.73	0.57	0.81	0.76	0.63	0.75	0.7

Conclusions

- Simple character n-grams and SVM can detect genres
- Linear SVM, 5 grams, $C = 10$ with feature selection
- GICR fully annotated to separate at least argumentative from newslike texts
- Issues with Personal and Promotion
- Full MDA for Russian to link text-external and text-internal descriptions
- Neural networks: promising, but problems with interpretation

References

Adamzik, K. (1995).
Textsorten – Texttypologie. Eine kommentierte Bibliographie.

Baroni, M. and Bernardini, S. (2006).
A new approach to the study of translationese.
Literary and Linguistic Computing, 21(3):259–274.

Biber, D. (1988).
Variations Across Speech and Writing.

Sharoff, S. (2018).
Functional text dimensions for the annotation of Web corpora.
Corpora, 13(1):65–95.

Sharoff, S., Wu, Z., and Markert, K. (2010).
The Web library of Babel: evaluating genre collections.
In *Proc LREC*, Malta.