Genres, registers and text functions

Stylistic properties of genres

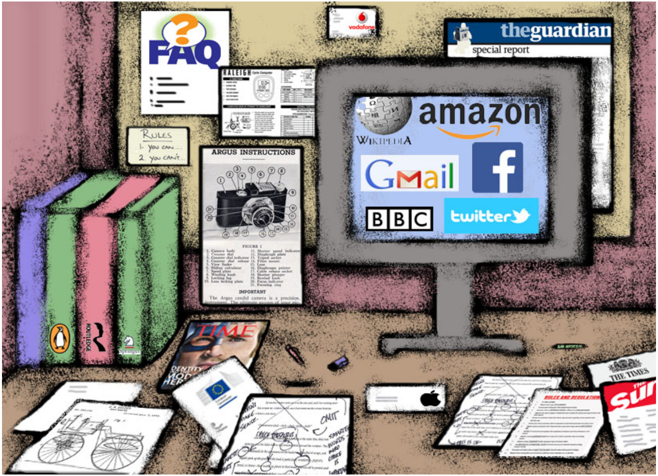
Serge Sharoff

Centre for Translation Studies
University of Leeds

9 January 2023

Outline

Genre topography on the Web scale



Saying sensible and useful things about any text (Halliday, 1985)

Language looks different when you look at a lot of it at once (Sinclair, 1991)

Problems:
1000s of genres
Hybridism

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

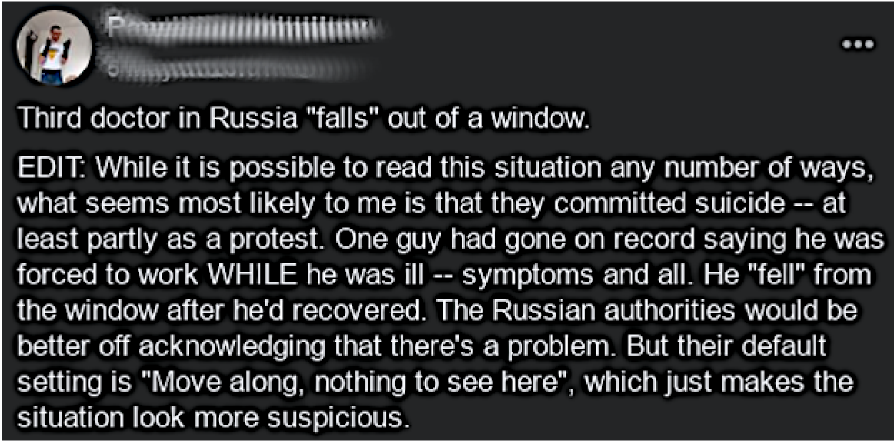
Hybridisation of genres

Brown 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 . . .

Reportage (A) or Editorial (B)?

The most positive element to emerge from the Oslo meeting of North Atlantic Treaty Organization Foreign Ministers has been the freer, franker, and wider discussions, animated by much better mutual understanding than in past meetings. This has been a working session of an organization that, by its very nature, can only proceed along its route step by step and without dramatic changes. . .
("NATO Welds Unity" The Christian Science Monitor, 1961)

Hybridisation and democratisation


Third doctor in Russia "falls" out of a window.

EDIT: While it is possible to read this situation any number of ways, what seems most likely to me is that they committed suicide -- at least partly as a protest. One guy had gone on record saying he was forced to work WHILE he was ill -- symptoms and all. He "fell" from the window after he'd recovered. The Russian authorities would be better off acknowledging that there's a problem. But their default setting is "Move along, nothing to see here", which just makes the situation look more suspicious.

Functional Text Dimensions (Sharoff, 2021)

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

Rating for functions with respect to prototypes

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

Functional Text Dimensions (Sharoff, 2021)

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

argum To what extent does the text try to persuade the reader? (*argumentative blogs* or *opinion columns*)

Rating for functions with respect to prototypes

0	none;	Ignore hesitations
0	slightly;	$\Uparrow$
.5	somewhat or partly;	$\Downarrow$
1	strongly.	Emphasise confident judgements

Functional Text Dimensions (Sharoff, 2021)

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

argum To what extent does the text try to persuade the reader? (*argumentative blogs* or *opinion columns*)

review To what extent does the text evaluate a specific entity? (*reviews of products or locations*)

Rating for functions with respect to prototypes

0	none;	Ignore hesitations
0	slightly;	$\Uparrow$
.5	somewhat or partly;	$\Downarrow$
1	strongly.	Emphasise confident judgements

Functional Text Dimensions (Sharoff, 2021)

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

argum To what extent does the text try to persuade the reader? (*argumentative blogs* or *opinion columns*)

review To what extent does the text evaluate a specific entity? (*reviews of products or locations*)

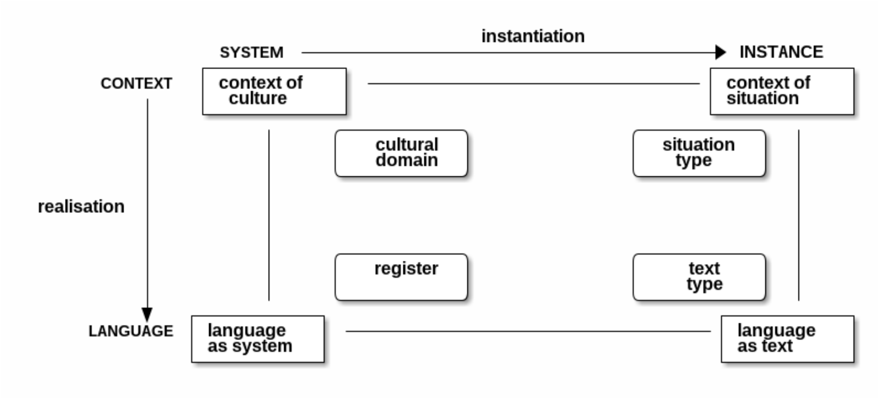
personal To what extent does the text report a personal story? (*diary-like blog entries*)

Rating for functions with respect to prototypes

0	none;	Ignore hesitations
0	slightly;	$\Uparrow$
.5	somewhat or partly;	$\Downarrow$
1	strongly.	Emphasise confident judgements

Form vs function in register classification

- External and internal classification criteria for Sinclair
- Texts in communication for Halliday



Principal Functional Text Dimensions

Code	Label	Prototypes	En	Ru
A1	argum	Argumentative blogs or opinion pieces	375	345
A4	fiction	Fiction, myths, film plots	103	97
A7	instruct	Tutorials or FAQs	221	96
A8	news	Reporting newswire articles	207	538
A9	legal	Laws, contracts, copyrights, T&Cs	95	105
A11	personal	Diary-like blog entries	161	284
A12	promotion	Adverts	350	331
A14	academic	Academic research papers	126	223
A16	info	Encyclopedic articles or textbooks	244	313
A17	review	Reviews of products or experiences	102	257
		Total training texts	1562	1930
A13	propaganda	Non-commercial promotion	73	62
A20	appell	Small ads, requests, CFPs	69	31

Prediction of text positions in this space and the nearest prototype.

Pre-trained transformers for prediction

Lots of methods → Fine-tuning XLM-Roberta

```
from transformers import pipeline
classifier = pipeline("text-classification",model="ssharoff/genres")
print(classifier("Never in the field of human conflict was so much owed
[{'label': 'A1', 'score': 0.842}, {'label': 'A8', 'score': 0.081}]
print(classifier("The most positive element to emerge from the Oslo mee
[{'label': 'A1', 'score': 0.328}, {'label': 'A8', 'score': 0.181}]
print(classifier("Alice was beginning to get very tired of sitting by h
[{'label': 'A4', 'score': 0.776}, {'label': 'A16', 'score': 0.071}]
print(classifier("Алисе наскучило сидеть с сестрой без дела на берегу р
[{'label': 'A4', 'score': 0.562}, {'label': 'A16', 'score': 0.082}]
print(classifier("Аліса тяжко занудьгувала, сидячи на березі без діла.
[{'label': 'A4', 'score': 0.452}, {'label': 'A16', 'score': 0.068}]
```

Zero-shot multilingual transfer to Ukrainian

Classification accuracy

FTD	F1		CI
Argument	0.729	±	0.021
News	0.944	±	0.011
Review	0.711	±	0.030
Personal	0.725	±	0.028
Promotion	0.937	±	0.012
Academic	0.883	±	0.023
Information	0.657	±	0.047
Instruction	0.760	±	0.104
Legal	0.757	±	0.039
Fiction	0.690	±	0.051

Confidence intervals from 10-fold cross-validation
Overall macro-F1 is 0.78± 0.037

Confusion matrix for classification

Predicted→ Reference↓	A1	A4	A7	A8	A9	A11	A12	A14	A16	A17
A1.argument	187	3	3	9	3	15	3	4	11	13
A4.fiction	6	50	0	0	0	12	0	0	2	6
A7.instruct	4	0	41	1	1	3	7	4	6	2
A8.news	13	0	1	446	2	3	2	1	8	4
A9.legal	4	0	0	0	59	2	1	2	5	0
A11.personal	13	3	0	2	0	134	3	0	4	18
A12.promotion	3	0	0	3	0	8	276	5	6	7
A14.academic	9	0	1	0	3	4	2	161	6	0
A16.info	19	0	0	11	4	10	1	10	90	6
A17.review	6	0	0	4	0	9	1	0	3	136

Composition of large Web corpora for pre-training

FTD		Wiki		OWT		CC-en		CC-ru
argum	*0.88%*	30720	21.21%	549016	17.08%	28735602	*10.05%*	787898
fiction	0.05%	1677	0.26%	6660	0.46%	771610	0.10%	8196
instruc	0.30%	10509	3.76%	97298	4.76%	8013691	3.12%	244983
news	1.14%	39665	49.78%	1288525	15.88%	26716672	26.84%	2104093
legal	0.04%	1340	0.11%	2731	1.90%	3190328	6.43%	504067
person	0.03%	1168	4.48%	116078	9.35%	15727798	*1.42%*	111731
promote	0.07%	2390	10.99%	284365	29.31%	49306151	27.45%	2152335
academic	0.82%	28558	0.72%	18720	1.77%	2983096	3.46%	271039
info	91.98%	3196502	*2.22%*	57342	11.64%	19590438	18.36%	1439570
review	4.68%	162511	6.48%	167772	7.85%	13209782	2.77%	216904

Pre-trained transformer models (BERT, Roberta, GPT):

- BERT for English is trained by combining Wiki and fiction, mBERT is trained on Wikipedia for all languages
- OWT (used in GPT-2) comes from upvoted links on Reddit
- CC (used in XLM-Roberta) comes from Common Crawl

COVID message journey: masks not needed

Research papers → policy making → mass media → social media

2010-09 The Annals of Occupational Hygiene *masks tested in the study had 40–90% instantaneous penetration levels against polydisperse NaCl aerosols employed in the respirator test protocol at 5.5 cm/s. Results show that common fabric materials may provide marginal protection against nanoparticle-sized viruses.* → `Function: academic, Face masks: not needed`

2020-04-20 NY Times *The C.D.C. has recommended that all Americans wear cloth masks if they go out in public.* → `Function: instruction, Region: USA, Face masks: support`

2020-04-30 Gov.UK News *Today, we're publishing our PPE plan: being clear who needs PPE, when they need it and who does not. . . . the best way to protect yourself and to protect others is to regularly wash your hands. And of course the most important way to protect yourself is to stay at home. Because a front door is better than any face mask.* → `Function: instruction, Region: UK, Face masks: not needed`

Message journey: support for wearing masks

2020-02-09, The Daily Telegraph *Increasing numbers of commuters on the London Underground are now wearing face masks, twitchily keeping watch on the respiratory condition of those around them.* → `Function: news, Region: UK, Face masks: support`

2020-02-11, Twitter *I told yall I'm anti vaxx but I'm not stupid like the ppl in the world. I'm the type who wears face masks when I'm in public.* → `Function: argument, Region: USA, Face masks: support`

2020-05-03 Reddit *With the lockdown being extended indefinitely, I'd like to have myself personalised face masks, something more stylish and durable. Current set of disposable masks don't last very long. Looking to support someone local to South London.* → `Function: argument, Region: UK, Face masks: support`

Corpus collection via crawling

	Source	Time span	#Texts	#Words	AWL
Research papers:	CORD19	1922–2021	183,185	678,679,071	3,705
Authority sources:	CDC FAQ	2020-08	645	29,375	46
	News from gov.uk	2020-01–2021-03	1,576	1,165,237	739
Newspapers&Policy	The Sun	2020-01–2021-03	21,704	14,724,045	678
	The Telegraph	2020-01–2021-03	16,510	20,991,551	970
	The Guardian	2020-01–2021-03	28,766	20,825,021	724
	World Economic Forum	2020-01–2021-03	2,528	2,440,099	965
Social Media:	Reddit	2020-02–2020-07	107,973	7,711,589	71
	Twitter	2019-12–2021-01	30,016,828	688,073,844	23

Botometer for filtering real users in Twitter

Extraction of texts on topic and genre

- Using Embedded Topic Models (Dieng et al., 2020)
  Topic6: *masks, social, distancing, face, wear, hands, wearing, air, water, food, spread, wash, ppe, sanitizer, buy, protect*

FTD	Reddit	Twitter	Guardian	Sun	Telegraph
Argument	56.43%	54.72%	39.30%	34.37%	41.97%
News	1.53%	11.16%	49.27%	49.86%	44.06%
Personal	22.14%	10.83%	4.59%	2.41%	5.50%
Promotion	1.37%	11.40%	4.59%	8.73%	5.04%
Academic	0.02%	0.03%	0.02%	0.02%	0.08%
Information	1.66%	1.02%	0.58%	1.32%	0.94%
Review	7.19%	7.20%	0.36%	0.06%	0.26%
Fiction	0.11%	0.21%	0.25%	0.08%	0.42%
Instruction	9.42%	3.29%	1.05%	3.13%	1.74%
Legal	0.13%	0.15%	0.00%	0.01%	0.01%

Misinformation potential for 'reporting personal experiences' and Instructions and Reviews

Explaining neural networks through linguistics

- Large corpus $\rightarrow$ composition
  Machine Learning predictions $\rightarrow$ reasons

Explaining neural networks through linguistics

- Large corpus $\rightarrow$ composition
  Machine Learning predictions $\rightarrow$ reasons
- Limits of interpretation via individual words and texts

Explaining neural networks through linguistics

- Large corpus $\rightarrow$ composition
  Machine Learning predictions $\rightarrow$ reasons
- Limits of interpretation via individual words and texts
- Lexicogrammatical categories from (Biber, 1988):

Explaining neural networks through linguistics

- Large corpus → composition
  Machine Learning predictions → reasons
- Limits of interpretation via individual words and texts
- Lexicogrammatical categories from (Biber, 1988):
  - Lexical (publicVerbs, timeAdverbials, amplifiers, ...)

Explaining neural networks through linguistics

- Large corpus $\rightarrow$ composition
  Machine Learning predictions $\rightarrow$ reasons
- Limits of interpretation via individual words and texts
- Lexicogrammatical categories from (Biber, 1988):
  - Lexical (publicVerbs, timeAdverbials, amplifiers, ...)
  - POS (prepositions, past tense verbs, nominalisations, ...)

Explaining neural networks through linguistics

- Large corpus → composition
  Machine Learning predictions → reasons
- Limits of interpretation via individual words and texts
- Lexicogrammatical categories from (Biber, 1988):
  - Lexical (publicVerbs, timeAdverbials, amplifiers, ...)
  - POS (prepositions, past tense verbs, nominalisations, ...)
  - Syntactic (*be* as the main verb, by-passives, ...)

Explaining neural networks through linguistics

- Large corpus → composition
  Machine Learning predictions → reasons
- Limits of interpretation via individual words and texts
- Lexicogrammatical categories from (Biber, 1988):
  - Lexical (publicVerbs, timeAdverbials, amplifiers, . . .)
  - POS (prepositions, past tense verbs, nominalisations, . . .)
  - Syntactic (*be* as the main verb, by-passives, . . .)
  - Text-level categories (type-token ratio, word length, . . .)

Explaining neural networks through linguistics

- Large corpus → composition
  Machine Learning predictions → reasons
- Limits of interpretation via individual words and texts
- Lexicogrammatical categories from (Biber, 1988):
  - Lexical (publicVerbs, timeAdverbials, amplifiers, ...)
  - POS (prepositions, past tense verbs, nominalisations, ...)
  - Syntactic (*be* as the main verb, by-passives, ...)
  - Text-level categories (type-token ratio, word length, ...)
- Ideally lexicogrammar features from (Halliday, 1985)

Explaining neural networks through linguistics

- Large corpus → composition
  Machine Learning predictions → reasons
- Limits of interpretation via individual words and texts
- Lexicogrammatical categories from (Biber, 1988):
  - Lexical (publicVerbs, timeAdverbials, amplifiers, . . .)
  - POS (prepositions, past tense verbs, nominalisations, . . .)
  - Syntactic (*be* as the main verb, by-passives, . . .)
  - Text-level categories (type-token ratio, word length, . . .)
- Ideally lexicogrammar features from (Halliday, 1985)
- Relative importance by training logistic regression

$$ln\frac{p}{1-p} = w_0 + w_1 x_1 + ... + w_n x_n$$

Variation in features across text functions

Rates of nominalisations (E14), nouns (E16), *by*-passives (F18), public verbs (K55) and clause negation (P67)

	E14		E16		F18		K55		P67	
	mean	median	mean	median	mean	median	mean	median	mean	median
Overall:	2.92%	2.46%	19.17%	18.96%	0.10%	0.00%	0.24%	0.00%	12.54%	7.30%
Argum	3.29%	2.99%	17.90%	17.75%	0.10%	0.00%	0.36%	0.27%	17.58%	12.99%
Fiction	1.38%	1.19%	14.77%	14.57%	0.07%	0.00%	0.59%	0.46%	26.32%	17.10%
Instruct	2.73%	2.32%	19.48%	19.36%	0.08%	0.00%	0.21%	0.00%	16.04%	10.69%
News	3.20%	2.97%	18.39%	18.18%	0.13%	0.00%	0.55%	0.43%	9.11%	4.16%
Legal	5.36%	5.16%	19.65%	19.56%	0.18%	0.12%	0.29%	0.19%	21.82%	15.75%
Personal	1.66%	1.39%	16.73%	16.56%	0.06%	0.00%	0.33%	0.23%	13.99%	9.92%
Promote	3.42%	3.03%	21.03%	20.95%	0.08%	0.00%	0.14%	0.00%	8.60%	0.00%
Academ	4.28%	3.99%	20.39%	20.35%	0.13%	0.00%	0.17%	0.03%	8.90%	0.00%
Info	2.50%	2.07%	17.87%	17.66%	0.15%	0.00%	0.15%	0.00%	8.11%	0.00%
Review	1.76%	1.59%	17.56%	17.50%	0.07%	0.00%	0.26%	0.14%	13.15%	9.38%

Variation in Russian

Rates of nominalisations (E14), nouns (E16), *by*-passives (F18), public verbs (K55) and clause negation (P67)

	E14		E16		F18		K55		P67	
	mean	median	mean	median	mean	median	mean	median	mean	median
Overall:	6.05%	5.46%	21.43%	21.42%	0.28%	0.15%	0.14%	0.00%	8.99%	7.68%
Argum	6.05%	5.47%	19.44%	19.41%	0.23%	0.15%	0.18%	0.10%	12.49%	11.73%
Fiction	2.42%	2.18%	18.32%	18.21%	0.21%	0.06%	0.27%	0.16%	15.47%	14.78%
Instruct	4.87%	4.48%	21.12%	21.11%	0.22%	0.11%	0.12%	0.00%	12.46%	11.60%
News	6.61%	6.16%	22.31%	22.22%	0.26%	0.00%	0.26%	0.00%	5.30%	3.38%
Legal	11.40%	11.07%	22.50%	22.41%	0.69%	0.59%	0.11%	0.00%	6.46%	5.04%
Personal	3.07%	2.80%	18.39%	18.19%	0.15%	0.00%	0.18%	0.07%	14.01%	13.33%
Promote	6.88%	6.46%	22.77%	22.64%	0.32%	0.17%	0.09%	0.00%	6.43%	5.13%
Academ	10.17%	9.80%	22.96%	22.93%	0.39%	0.31%	0.10%	0.00%	5.42%	4.29%
Info	6.32%	5.78%	22.25%	22.27%	0.35%	0.27%	0.09%	0.00%	6.87%	5.96%
Review	3.90%	3.59%	19.49%	19.39%	0.22%	0.00%	0.14%	0.00%	11.78%	10.81%

Text difficulty for easy instructions and news

A7.instructional		A8.news	
C07.2persProns	0.5155	K55.publicVerbs	0.2913
C06.1persProns	0.1791	H35.causative	0.2666
B04.placeAdverbials	0.1702	H38.otherSubord	0.2214
I39.preposn	0.1603	N59.contractions	0.2192
L54.predicModals	0.1371	K47.generalHedges	0.2129
N60.thatDeletion	0.1341	D13.whQuestions	0.1841
B05.timeAdverbials	0.1028	A01.pastVerbs	0.1756
L53.necessModals	0.0638	C09.impersProns	0.1525
H35.causative	-0.0784	C08.3persProns	0.0521
K56.privateVerbs	-0.0902	F18.BYpassives	-0.1857
H25.presPartClaus	-0.0984	K48.amplifiers	-0.1864
E14.nominalizations	-0.1146	K50.discoursePart	-0.2290
I42.ADV	-0.1366	L54.predicModals	-0.2427
C09.impersProns	-0.1612	E16.Nouns	-0.2705
A03.presVerbs	-0.1678	K45.conjuncts	-0.3521
E16.Nouns	-0.2482	C07.2persProns	-0.4385

Take-home message

- Genre typology vs topology
  Functional Text Dimensions with respect to prototypes
- You should always compare your findings across corpora and across genres
  Differences across popular corpora for pre-trained models
- Differences in COVID genre classification
- Compare text-external functions using text-internal features
  Also across languages: nominalizations or predictionModals

Available resources

Classifier `https://huggingface.co/ssharoff/genres`

Biber features `https://github.com/ssharoff/biberpy`

References I

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

Biber, D. (1988).
Variation Across Speech and Writing.
Cambridge University Press.

Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2020).
Topic modeling in embedding spaces.
Transactions of the Association for Computational Linguistics, 8:439–453.

Halliday, M. A. K. (1985).
An Introduction to Functional Grammar.
Edward Arnold, London.

References II

Sharoff, S. (2021).
Genre annotation for the web: text-external and text-internal perspectives.
Register studies, 3:1–32.