Lexical variation in English language podcasts, editorial media, and social media

Jussi Karlgren, Spotify, Stockholm, Sweden

Abstract The study presented in this paper demonstrates how transcribed podcast material differs with respect to lexical content from other collections of English language data: editorial text, social media, both long form and microblogs, dialogue from movie scripts, and transcribed phone conversations. Most of the recorded differences are as might be expected, reflecting known or assumed difference between spoken and written language, between dialogue and soliloquy, and between scripted formal and unscripted informal language use. Most notably, podcast material, compared to the hitherto typical training sets from editorial media, is characterised by being in the present tense, and with a much higher incidence of pronouns, interjections, and negations. These characteristics are, unsurprisingly, largely shared with social media texts. Where podcast material differs from social media material is in its attitudinal content, with many more amplifiers and much less negative attitude than in blog texts. This variation, besides being of philological interest, has ramifications for computational work. Information access for material which is not primarily topical should be designed to be sensitive to such variation that defines the data set itself and discriminates items within it. In general, training sets for language models are a non-trivial parameter which are likely to show effects both expected and unexpected when applied to data from other sources and the characteristics and provenance of data used to train a model should be listed on the label as a minimal form of downstream consumer protection.

1 Genres and podcast transcripts

The way human language is used varies across channels and styles, and we have for the longest while made a clear distinction between spoken and written language as two major distinctive modes of communication (Cederschiöld, 1897; Ong, 1982; Biber, 1991; Coulmas, 2003).

The differences between writing and reading can to a large extent be related to situational differences: where speech has been used in transient situations in which interlocutors are present, writing has typically been used in asynchronous communication with participants at a remove from each other. This distinction has through the introduction of communication technologies become less and less clear-cut. Written language is used for momentary and fleeting conversations with little planning or editorial oversight; spoken language material is created, published and distributed in ways which are more formal and more permanent and archival than before.

Podcasts are a new medium and a new format for spoken language. The styles of language use in podcasts are as yet unformed and have not yet coalesced into stable functional and generally accepted genres: podcast material will require us to recalibrate many of the assumptions we make about how language is used. Recently, a collection of over 100,000 podcast episodes, including automatically generated transcripts, has been released for the purposes of retrieval and summarisation experimentation. The companion paper released with the podcast material set gives some indicative differences between the transcripts and written language as represented by the Brown corpus (Francis and Kucera, 1967) and shows i.a. that the frequency of amplifiers and personal pronouns is greater than in the various genres represented in the Brown corpus (Clifton et al., 2020).

This paper demonstrates how some such differences across text collections of different types are indicative of genre differences, some of which can be expected to depend on how spoken genres continue to evolve with changing technology and evolving situations of usage. This examines differences in anchoring, subjective language, and discourse handling, which can all be expected to be dimensions in which podcast language will differ from written genres.

Podcasts are a rapidly evolving medium. The variation and volatility is great and we can expect that only a few years from now there will be new formats of language use not represented in the present collection. These measurements are intended to inspire the systematic exploration of such differences as they occur, to make possible documentation of current and future

changes in the medium, to make explicit differences that may have effects on the applicability of language models trained on one type of material on another, and to ensure that application to classification, retrieval, or large scale extraction of information is informed and sensitive to those systematic differences that might impact results.

2 Data Overview

Seven data sets were used for these experiments. These data sets are of varying age and collected with various methods, but have all been used in research and benchmarking projects and are selected by virtue of being accessible for experimentation and further study. The representativeness of the corpora may vary: movie scripts change over time as the craft of writing and acting evolves; conventions in phone conversations change as new technologies cater to new use cases; social media platforms, with various conventions and various technological affordances, go in and out of fashion; editorial media shift their focus and their offerings according to the shifting constitution and preferences of their audiences. The editorial media data set is the one most clearly governed by conventions and constraints imposed by audience expectations for the genre and is likely to be the data set with least change over time. These changes are all likely to affect the stylistic statistics on reported below in various ways; the differences found between genres are robust enough to majorise the within genre differences over time.

Editorial media A collection of Associated Press newswire text from year 1989-1990 made available for experimentation in various shared tasks as part of the TIPSTER corpus (Harman and Liberman, 1993). These represent edited text conforming to standard written English language usage.

> *... Citing financial disarray in Massachusetts government, a major bond rating agency cut the state's credit rating Friday for the second time this year, a move that could add millions to borrowing costs. The decision by Standard & Poor's Corp. to downgrade Massachusetts bonds from AA-to A represents a harsh assessment of the fiscal policies of Gov. Michael S. Dukakis and the state Legislature. "The state's economy remains strong, while debt and fiscal management display serious weaknesses," the agency said. ...*

Social media Data from the Blog Authorship Corpus which consists of a large age- and gender-balanced collection of about 700 000 English language blog posts collected for the purpose of experimentation with authorship attribution (Schler et al., 2006). These are intended to represent informal written language in a variety of subgenres.

> *... Yesterday I learned a new programming language, Groovy . Well, I wrote a simple program in Groovy. I need to do much more with it before I learn to "think in Groovy." This is important. There's a huge benefit to learning a new programming language, so much so that The Programatic Programmers recommend learning a new language every year. Learning a new programming language can be difficult. Let's be precise: learning to write working programs in a new language is relatively easy, but the first impulse is to think in the style of the languages you already know and write programs using the syntax of the new language. ...*

Microblogs A set of mostly English language microblog posts from Twitter collected for analysis of public opinion during the fall of 2017. [1]. These are intended to represent real-time language use, but in fact contain a large number of press release announcements, news headlines, and links to further reading.

- *Mystery Fanged Sea Creature Washes Up on Texas Beach after Hurricane Harvey* `URL`
- *Hope and kudos for hurricane victims in healthcare:* `URL`
- *as i sit in this heat i also wish tha best for those that caught harvey cause i know theyre worse off n im grateful we ain get hit directly*
- *Having a gun license is what you're thinking about after a disaster? If you're in Taaaxas. #Harvey* `URL`
- `@UId` *Hi, sorry missed your question! 7-8pm at harvey hadden, this will be the only one i'm afraid*

Podcast transcripts A large collection of automatically generated English language podcast transcripts released by Spotify for research purposes, with episodes representing a variety of podcast formats, styles, levels of formality, and topics (Clifton et al., 2020). The transcripts include sentence breaks automatically inferred by the transcription system.

> *... Only on my hands no with my hips ever. So first what I did was visiting a doctor because every time when I was trying to stretch*

[1] The post ids are available at `http://www.lingvi.st/corpora/storm.txt`

myself like to take stretch classes, I ended up with like a really bad pain for like a few weeks or months. So then they visited doctor and I really like he told me that my spine like ...

Movie scripts A collection of English language movie scripts from the Film Corpus (Walker et al., 2012). The corpus has separated dialogue from scene descriptions and director instructions; for the purpose of this study, only the dialogue portion has been used, as a sample of language which is produced in written form but intended to represent natural speech.

...
- What's that shit?
- A book. It's called reading. You should try it some time.
- You wanna read something. Read between the lines.
- Well here's something even you can relate to. Albert got a lotta trim.
- That genius thing is a babe magnet.
- Lemme see that book.
...

Telephone conversations The Switchboard corpus is a collection of transcribed English language telephone conversations on a variety of topics (Godfrey et al., 1992; Godfrey and Holliman, 1997). For this study a separately annotated portion which is freely available is used (Jurafsky et al., 1997). This is intended to represent the character of spontaneous unscripted speech. This transcription is fairly carefully done to preserve e.g. interruptions and overlapping speech, in contrast with the podcast transcription.

...
— What kind of ...
— Okay.
— ... eating out do you enjoy?
— Well, I like dining out.
— Of course, it means that I don't have to cook.
— Right .
— But, um, I'm a divorced woman.
— I have one child ...
— Uh-huh.
— ... and, you know, when, when we dine out we go to like medium priced restaurants.
— Uh—huh.
— I don't, I don't particularly
— I think it's sort of a waste of money to go real, to a real high priced restaurant.
— Do you go like home cooking, like Black-Eyed Pea and that kind of thing or ...
— Um, e-, n-
— ... cafeteria?
— Not really.
— We go wh-, more for the, uh, Chinese ...
— Me too .
— ... and Italian ...
— Uh-huh.
— ... and stuff like that. Mexican, stuff ...
— Mexican,
— uh-huh.
— ... that I can't cook .
— Uh, we do too.
— We do the same.
— Yeah.
...

Popular lectures The popular science TED talk series on "technology, entertainment, and design" provide transcripts of lectures given by the speakers. The lectures are information-dense, but informal and entertaining in style and are mostly monologues, with the occasional conversational interview. A selection of such transcripts has been made available for experimentation (Banik, 2017).

I'd like to tell you the tale of one of my favorite projects. I think it's one of the most exciting that I'm working on, but I think it's also the simplest. It's a project that has the potential to make a huge impact around the world. It addresses one of the biggest health issues on the planet, the number one cause of death in children under five. Which is...? Water-borne diseases? Diarrhea? Malnutrition? No. It's breathing the smoke from indoor cooking fires — acute respiratory infections caused by this.

100 000 sentences from each source were sampled for inclusion in this study, using the Natural Language Toolkit (NLTK) for sentence segmentation which splits the text to sentences at major delimiters (".", "!", "?") and at paragraph breaks (Bird, 2006). Some quantitative data for the samples are given in Table 1. Noticeable is that the sentence length varies considerably across the collections. This reflects both genre variation and transcription practice, as can be seen in the above example extracts: the movie scripts contain very short sentences authored to describe rapid dialogue and overlapping turns, the phone conversation transcripts render short turns and interruptions as separate sentences, where, by contrast, the podcast transcripts have longer turns on average. Repeated samples were drawn to ensure stability of the measures made, and all measures and statistics given in the following tables are averaged

across a number of resamplings, rounded to two significant figures.

3 Dimensions of variation

The measures examined in this study focus on readily inspectable aspects of language use where spoken language and informal channels traditionally are assumed to show difference to written formal genres. Spoken language due to its immediacy and synchronous nature frequently has more overt markers for interpersonal functions, and utilises different textual functions to organise the discourse. Since written genres more frequently are used for abstract and complex topical matter, it is to be expected that those ideational functions that concern argumentation and logical structure are rendered differently. Biber and colleagues, in their studies on register variation across several languages (Biber, 1995), posit a number of variational dimensions using factorial analyses and then formulate a low dimensional space of *functional bases* in which they position the genre samples such as lectures, face-to-face conversations, broadcasts, private letters, academic prose, offical documents, and many more.

This study uses a subset of the variables examined by Biber and colleagues (the variables used by Biber are variously accessible for automated analysis). The addition of podcast material to the data used by Biber are likely to extend the variational dimensions posited by his original study, since podcast material cuts across many of the suggested dimensions such as "involved vs informational" which separates e.g. speeches from e.g. academic prose; "narrative vs non-narrative", which separates fiction text from e.g. face-to-face interaction; "textual vs situational reference", which separates e.g. phone conversations from official documents and so forth. Podcasts incorporate material with the situatedness of personal conversations to the abstraction of formal lectures, and material with the immediacy and interactive online planning of live dialogue to the editorially oriented production qualities of broadcast news. We can expect that many of the variational dimensions are relevant for podcasts even as new conventions and new genres gradually develop.

Spoken unscripted language is characterised by explicit features related to the organisation of discourse which involve turn-taking, interruptions, dysfluencies, and repair. These are somewhat challenging to study with the given collections, especially as transcription oftentimes removes and normalises much of the signal. Notably, in the present collections, while the phone conversation transcripts render turn-taking in detail, the podcast transcriptions leave out overlapping speech.

This study focusses on features of language use which is *situated*, where the participants are synchronously present during the communicative situation as opposed to communication where the author or speaker is separated from the audience, and *personal and subjective*, where the attitude and stance of the author or speaker is clearly expressed and modulated to capture the attention and fit the reactions of the intended audience, in contrast to language framed to be formal and couched in objective terms and expressions, abstracted from the present situation.

The surface features to be expected are more attitudinal and overtly subjective language, with intensifiers, first and second person pronouns, more present tense and narratives, more questions and affirmations than in scripted and planned language use.

4 Attitude and Affect in Language

Subjective language is of interest for many reasons, but not least for its potential applications in information retrieval and text categorisation. Since the introduction of computational sentiment analysis as a research topic (Qu et al., 2004) various efforts to extend or typologise the field have been explored, (Karlgren et al., 2004; Karlgren, 2009; Feldman, 2013; Ravi and Ravi, 2015) and many mostly lexical approaches were implemented for commercial application. Now, with computational methods that allow full scope over an entire utterance without relying on single items, some of the lexical approaches are less immediately impressive than before, but for reasons of transparency, many are still in use in practical applications and they correlate well with findings from non-lexical approaches. For the present experiments, a standard lexicon of polar items has been used to represent the manifold expressions of human emotion found in text (Hu and Liu, 2004), and the incidence of items from the lexicon are shown in Table 2[2].

The table gives counts both per word, i.e. how many of the tokens of the collection sample were polar evaluative lexical items (left half of the table), and per sentence, i.e. how many sentences of the 100 000 sample contained a polar evaluative lexical item (right half of the table).

The results show that podcast transcripts have a noticeably higher incidence of positive polarity items and lower incidence of negative polarity items than written genres and that popular lectures exhibit much the same distribution. News stories exhibit more negative polarity than positive polarity items, which is likely to

[2]The item "*like*" was removed from the list of positive items, since it is very frequent in the spoken language material as a non-attitudinal discourse particle.

Table 1: Descriptive statistics for the seven language collections, comparing average sentence length.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone calls	Popular lectures
Number of sentences	100 000	100 000	100 000	100 000	100 000	100 000	100 000
Number of words	2 200 000	1 700 000	1 800 000	1 700 000	720 000	720 000	1 600 000
Words per sentence	22	17	18	17	7.2	7.2	16
Year of publication	1989-1990	2004	2017	2019	before 2010	early 1990s	2017

have to do with editorial considerations: negative news drive reporting. This probably explains a similar imbalance for the microblog posts, which to a large extent are commentary to current news stories. The two more traditional spoken genres have much lower counts of both polarities, which may mean that the lexicon used here is not optimised for spoken material or that spoken language demonstrates polarity more often in constructional items rather than purely lexical ones ("*This put me off.*").

5 Amplification

Amplifiers are linguistic items that serve to increase the perceived strength of an evaluative expression. They are typically constructed as adverbials, as shown in Example (1) (Quirk et al., 1985, §7.57 a) and in this study such items are used, and other amplifying constructions are left aside. Amplifiers can be subcategorised in several ways, and here a three-way distinction is made. *Gradation* amplifiers increase the intensity of a gradal expression: (*very*, *immensely*, *substantially*, *fucking*); *affirmation* amplifiers emphasise the commitment of the speaker to the sentiment (*truly*, *really*); and *surprise* amplifiers communicate that the qualities under consideration are unexpected or anomalous (*amazingly*, *surprisingly*, *unusually*). These distinctions are of course not independent of each other. The amplifiers used in this study are given in Appendix A.

(1) a. Hurricane Irma is a ***very*** dangerous storm. (MICROBLOG)
b. The ***immensely*** popular "Star Wars" isn't much good for teaching science. (NEWS)
c. It just ***fucking*** cool. (PODCAST)
d. If you're ready to find out who you are deep down and live a ***truly*** authentic life. (PODCAST)
e. Now if you use the right kind of atoms and you get them cold enough, something ***truly*** bizarre happens. (LECTURES)
f. My husband is he's ***really*** sweet. (PODCAST)
g. Leaders of corporate America say business is ***surprisingly*** good. (NEWS)
h. That was interesting, and ***surprisingly*** nice. (BLOG)

The incidence of amplifiers in the seven collections are given in Table 3. We find that the podcast material has an order of magnitude higher number of amplifiers than most other genres. Popular lectures also exhibit a similarly high incidence of amplifiers, but there is a difference in how they are distributed over the subcategories: podcasts show a very high incidence of *affirmation* amplifiers, which take purchase in the presence of the speaker in the communicative situation. This is one of the most differentiating features between podcasts and popular lectures, which otherwise exhibit many similar characteristics.

6 Negation

Negation is a foundational semantic operator whose exact semantic function on the meaning of an utterance can be discussed and modelled at length (Von Klopp, 1993, e.g.). Negation can affect an entire clause ("*I didn't eat the cookies.*") or more locally, a constituent of a clause ("*I will eat no more cookies.*"). In English, clausal negation most often is formed through the negative verbal affix "*n't*", which in written or more formal registers, or when emphasised, often is rendered as the separate lexical item "not". Local negation is formed through prefixing the negated component with "*no*" or "*not*", or by using more elaborate construction such as "*neither ... nor*", "*nobody*", "*none*", or "*never*" (Quirk et al., 1985, §10.55ff). Negation has an obvious relation to polarity and antonymy which has motivated great interest in research on methods for the practical handling of negation in sentiment analysis and related experiments and applications (Choi and Cardie, 2009; Tanushi et al., 2013; Mohammad et al., 2013; Kiritchenko et al., 2014; Reitan et al., 2015, i.a.). Some examples of negation and its effect on polarity are given in Example (2). In this study, negation is included as an example of an accessible semantic operator useful for modulation and modification of attitudinal expressions. The list of negations used in this study, compiled from Quirk et al. (1985) and Biber (1995) is given in Appendix B and the incidence of negations is given in Table 4. We can here observe how informal genres, unsurprisingly, exhibit many more contracted forms than the written material. We also find that the incidence of

Table 2: Occurrence and proportion of negative and positive polar lexical items from Hu and Liu (2004) in seven collections of language, per word and per sentence in a sample of 100 000 sentences from each collection.

	Per word				Per sentence	
	Positive		Negative		Positive	Negative
Editorial media	39 000	(1.8 %)	56 000	(2.6 %)	30 000	39 000
Social media	44 000	(2.6 %)	41 000	(2.5 %)	31 000	28 000
Microblogs	27 000	(1.5 %)	42 000	(2.3 %)	21 000	29 000
Podcast transcripts	46 000	(2.7 %)	29 000	(1.7 %)	33 000	21 000
Movie scripts	16 000	(2.2 %)	18 000	(2.6 %)	14 000	16 000
Phone conversations	16 000	(2.3 %)	9 000	(1.3 %)	15 000	8 100
Popular lectures	41 000	(2.6 %)	29 000	(1.8 %)	31 000	22 000

Table 3: Occurrence and proportion of lexical amplifiers (listed in Appendix A) in seven collections of language, per word and per sentence in a sample of 100 000 sentences from each collection.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations	Popular lectures
	Per word						
amplifiers	3 400	8 100	1 800	13 000	2 400	5 200	11 500
gradation	2 100	3 200	840	4 400	1 400	1 500	5 800
	(0.097 %)	(0.19 %)	(0.046 %)	(0.26 %)	(0.20 %)	(0.20 %)	(0.37 %)
affirmation	710	4 200	480	7 300	800	3 500	4 100
	(0.033 %)	(0.26 %)	(0.026 %)	(0.43 %)	(0.11 %)	(0.49 %)	(0.26 %)
surprise	580	680	470	950	160	220	1 600
	(0.027 %)	(0.041 %)	(0.026 %)	(0.055 %)	(0.021 %)	(0.031 %)	(0.10 %)
	Per sentence						
gradation	2 000	3 000	820	3 900	1 400	1 300	5 100
affirmation	700	3 900	440	6 400	780	3 400	4 100
surprise	570	660	410	910	160	220	1 600

negation in general is higher in social media and podcasts than in the other material. There are many hypothetical explanations for this observation which need further study: a tentative explanation is that negation is at times used as a discourse marker ("*No, no, no, no, we can't do that.*" or even "*No, you are right.*")

(2) a. We would continue to pursue the accelerator technology, but at the moment it is ***not*** as mature as fission reactors. (NEWS)
b. And it's crazy how it's it's ***not*** crazy. (PODCAST)
c. My boat got hit by #IrmaHurricane the ranch is #flooding from #irma but #hankjr02 is following me on Twitter, so it ***can't*** be all bad. (MICROBLOG)
d. and, it's ***not*** very expensive that way. (PHONE)
e. And that is ***not*** bad at all. (PHONE)
f. Ladies and gentlemen, a picture is ***not*** worth a thousand words. In fact, we found some pictures that are worth 500 billion words. (LECTURES)

7 Interrogatives

The incidence of interrogative utterances, defined as sentences that end with a "?", differs across the collections as shown in Table 5. It is likely, here as in preceding statistics, that the results are influenced by conventions for transcription which vary across the spoken genres, but the podcast material which is the only automatically transcribed material shows a higher incidence of questions than some of the other genres, rather than the lower incidence which might be expected from transcription errors. The movie script collection stands out here, with every sixth sentence a question, reflecting the type of conversational to-and-fro characteristic of the genre.

8 Situatedness

Personal pronouns are used when the author or speaker and the audience have a shared understanding of the context they are in. First and second person pronouns are less prevalent in formal discourse and more prevalent in face-to-face conversation than in other situations; narrative discourse will show a higher propor-

Table 4: Occurrence and proportion of negated sentences in seven collections of language in a sample of 100 000 sentences from each collection(negations used are listed in Appendix B).

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations	Popular lectures
negations	17 000	24 000	6 800	28 000	1 900	12 000	17 000
"no", "not"	10 000	9 000	3 500	9 600	4 200	3 800	8 400
contractions	3 800	9 900	4 000	13 000	8 800	7 100	7 300
constructions	1 800	2 500	1 300	2 100	2 300	1 100	1 300

Table 5: Occurrence and proportion of interrogative sentences in seven collections of language in a sample of 100 000 sentences from each collection.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations	Popular lectures
Questions	730	7 100	2 800	7 600	17 000	3 700	8 000

tion of third person pronouns than non-narrative discourse. The expected differences are found in the data, as shown in Table 6. These counts include reflexives ("*myself*") and possessives ("*our*", "*ours*"); the third person counts do not include "it"; the second person counts include impersonal "you" as in "when you bake bread you usually add some kind of leavening". Notable is firstly (and unsurprisingly) the high incidence of personal pronouns in the spoken genres together with the social media texts compared to the two other genres. Secondly, notable is the large number of second person pronouns in podcast material, movie scripts, and phone conversations, reflecting the dyadic conversational format in many of them. Thirdly, the large number of 1st person plural pronouns in the popular lecture data, reflecting the genre-specific pattern of including the audience in an utterance ("*When we think about why we hear, we don't often think about the ability to hear an alarm or a siren, although clearly that's an important thing.*") . A final striking observation is the consistently low level of reference to feminine correlates across all collections.

Another measure of situatedness is the distribution of lexical categories over content words. Table 7 shows how verbs are less common and proper nouns are more common in editorial media and in microblogs compared to the other four genres. These counts are based on part of speech tagging as provided by the NLTK part of speech tagger (Bird, 2006). The difference is most likely related to news reporting being based on participating people, organisations, and locales. By contrast, the relative occurrence of verbs is higher in the spoken genres and in social media.

Shared across all genres except the news material is the preponderance of present tense in comparison with past tense as shown in Table 8. This is an indicator of narrative discourse, where language is used to describe something that preceded the communicative situation. The news genre is highly focussed on reporting past events and this is reflected in the tense representation. These counts are also based on the NLTK part of speech tagger, which provides separate labels for past tense verbal forms. Some sentences have mixed tense, subclauses with a tense different from the matrix clause, e.g., or other more complex verbal structure and are omitted from the table.

9 Concluding Observations

This initial study demonstrates some clear differences in lexical content between transcribed podcast material and other collections of language data: editorial text, social media, both long form and microblogs, dialogue from movie scripts, transcribed phone conversations, and popular lectures. Most of the recorded differences are as might be expected, reflecting known or assumed difference between spoken and written language, between dialogue and soliloquy, and between scripted formal and unscripted informal language use. Most notably, podcast material, compared to the hitherto typical training sets from editorial media, is characterised by being in the present tense, and with a much higher incidence of pronouns and negations. These characteristics are, unsurprisingly, largely shared with social media texts. Where podcast material differs from social media material is in its attitudinal content, with many more amplifiers and much less negative attitude than in blog texts. There is a solid base to explain these differences in the studies by Biber referred to above, and in the more general notion of metafunctions of language which are utilised with various relative strength across communicative situations.

It is to be expected that the results presented in this study will age rapidly with respect to their details: the podcast medium will evolve and new genres and stylistic conventions will emerge an coalesce in the near future as podcasts gain a broader audience, more creators, and further situations of use. The popular lec-

Table 6: Occurrence of personal pronouns and their proportion of the vocabulary in seven collections of language.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations	Popular lectures
1 p sg	8 000	96 000	6 000	77 000	33 000	36 000	40 000
	(0.37 %)	(5.8 %)	(0.33 %)	(4.5 %)	(4.6 %)	(5.0 %)	(2.5 %)
2 p	2 500	16 000	7 200	52 000	23 000	18 000	30 000
	(0.11 %)	(0.96 %)	(0.39 %)	(3.0 %)	(3.2 %)	(2.6 %)	(1.9 %)
3 p sg m	22 000	14 000	4 700	14 000	9 000	3 100	7 600
	(1.0 %)	(0.82 %)	(0.26 %)	(0.81 %)	(1.2 %)	(0.44 %)	(0.48 %)
3 p sg f	4 500	8 500	2 000	5 900	4 100	2 000	3 700
	(0.21 %)	(0.51 %)	(0.11 %)	(0.34 %)	(0.56 %)	(0.29 %)	(0.23 %)
1 p pl	5 500	12 000	7 000	18 000	5 500	8 000	32 000
	(0.25 %)	(0.74 %)	(0.38 %)	(1.0 %)	(0.76 %)	(1.1 %)	(2.0 %)
3 p pl	11 000	7 000	4 800	13 000	2 600	9 200	15 000
	(0.50 %)	(0.42 %)	(0.26 %)	(0.74 %)	(0.36 %)	(1.3 %)	(0.97 %)

Table 7: Distribution of lexical categories for content words and their proportion of the vocabulary in seven collections of language based on NLTK part of speech tagging.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations	Popular lectures
verbs	350 000	320 000	230 000	370 000	150 000	140 000	300 000
	(16 %)	(19 %)	(12 %)	(21 %)	(21 %)	(20 %)	(19 %)
nouns	470 000	310 000	400 000	250 000	110 000	99 000	220 000
	(22 %)	(19 %)	(22 %)	(14 %)	(15 %)	(14 %)	(13 %)
proper nouns	260 000	92 000	460 000	63 000	52 000	21 000	51 000
	(12 %)	(5.5 %)	(25 %)	(3.7 %)	(7.2 %)	(2.9 %)	(3.1 %)
adjectives	160 000	120 000	170 000	96 000	36 000	39 000	110 000
	(7.4 %)	(7.1 %)	(9.2 %)	(5.6 %)	(5.0 %)	(5.4 %)	(6.7 %)

Table 8: Verb tense of sentences in seven collections of language in a sample of 100 000 sentences from each collection. Sentences with mixed tense or complex verb chains omitted.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations	Popular lectures
present tense	14 000	32 000	25 000	41 000	39 000	35 000	45 000
past tense	54 000	26 000	25 000	14 000	13 000	11 000	21 000

tures, an offshoot from classical academic lectures, but with their form modified by new transmission channels and by influence from other staged presentations, shows one direction of development which is clearly related to podcasts; we should expect some podcasts to adhere to this genre, while others will be more like drama and scripted speech, and some continue to exhibit similarities to more unscripted and informal conversation. Across all genres, the difference between *he* and *she* in their various forms was dramatic — this is something that may change over time.

These observations have some direct ramifications for computational work. Firstly, any useful approach to information access for material which is not primarily topical should be designed to be sensitive to such variation that defines the data set itself and discriminates items within it. More generally, training sets for language models are a non-trivial parameter which are likely to show effects both expected and unexpected when applied to data from other sources. The characteristics and provenance of data used to train a model should be listed on the label as a minimal form of downstream consumer protection. What these counts specifically demonstrate is that filtering the a data set through application of "stoplists" or other feature reduction methods or assessing the quality of language models using gold standards built on referential semantics based on nouns (cf. Karlgren (2019)) will reduce the richness of expression more in pronoun-rich and verb-rich genres than in those with less pronouns and verbs.

The variation demonstrated by the lexical tables given here is of obvious philological interest, casting light on how human communicative behaviour is modulated by the channel over which it proceeds. These reported statistics are but a scratch on the surface: more sophisticated and hypothesis-driven methods will be able to present more unified underlying variables and models with more explanatory power.

Acknowledgments

The author wishes to thank his colleagues Rosie Jones and Sravana Reddy for insightful comments which have contributed greatly to the quality of the paper.

References

Banik, Rounak. 2017. TED Talks: Data about TED Talks. *Dataset on Kaggle*, Version 3.

Biber, Douglas. 1991. *Variation across speech and writing*. Cambridge University Press, Cambridge.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Bird, Steven. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72. International Committee for Computational Linguistics.

du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000-2005. *Santa Barbara corpus of spoken American English*. Linguistic Data Consortium, Philadelphia.

Cederschiöld, Gustaf. 1897. *Om svenskan som skriftspråk [Swedish as a written language]*. Wettergren & Kerber, Gothenburg.

Choi, Yejin and Claire Cardie. 2009. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 590–598. Association for Computational Linguistics.

Clifton, Ann, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, et al. 2020. 100,000 podcasts: A spoken english document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics (Coling)*. International Committee for Computational Linguistics.

Coulmas, Florian. 2003. *Writing systems: An introduction to their linguistic analysis*. Cambridge University Press.

Feldman, Ronen. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

Francis, W Nelson and Henry Kucera. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence.

Godfrey, John J and Edward Holliman. 1997. Switchboard-1 Release 2. *Linguistic Data Consortium, Philadelphia*, 926.

Godfrey, John J, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 517–520. IEEE.

Harman, Donna and Mark Liberman. 1993. *TIPSTER Complete LDC93T3A Web Download*. Linguistic Data Consortium, Philadelphia.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International conference on Knowledge discovery and data mining (KDD)*, pages 168–177. Association for Computing Machinery.

Jurafsky, Dan, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL labeling project coder's manual, Draft 13. *Boulder Institute of Cognitive Science Technical Report*.

Karlgren, Jussi. 2009. Affect, appeal, and sentiment as factors influencing interaction with multimedia information. In *Theseus ImageCLEF workshop on visual information retrieval evaluation*, pages 8–11.

Karlgren, Jussi. 2019. How lexical gold standards have effects on the usefulness of text analysis tools for digital scholarship. In *International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF)*, pages 178–184. The CLEF Initiative.

Karlgren, Jussi, Gunnar Eriksson, Stefano Mizzaro, Paul Clough, Mark Sanderson, Kristofer Franzén, and Preben Hansen. 2004. Reading Between the Lines: Attitudinal expressions in text. In *Proceedings of the AAAI Spring Symposium Workshop on Exploring Attitude and Affect in Text: Theories and Applications*. American Association for Artificial Intelligence.

Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, pages 723–762.

Mohammad, Saif M, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, pages 321–327.

Ong, Walter J. 1982. *Orality and literacy*. Routledge.

Qu, Yan, James Shanahan, and Janyce Wiebe. 2004. *Proceedings of the AAAI Spring Symposium Workshop on Exploring Attitude and Affect in Text: Theories and Applications*. American Association for Artificial Intelligence.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Longman, London.

Ravi, Kumar and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.

Reitan, Johan, Jørgen Faret, Björn Gambäck, and Lars Bungum. 2015. Negation scope detection for twitter sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 99–108.

Schler, Jonathan, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of the AAAI Spring Symposium Workshop on Computational approaches to analyzing weblogs*, pages 199–205. American Association for Artificial Intelligence.

Tanushi, Hideyuki, Hercules Dalianis, Martin Duneld, Maria Kvist, Maria Skeppstedt, and Sumithra Velupillai. 2013. Negation scope delimitation in clinical text using three approaches: NegEx, PyConTextNLP, and SynNeg. In *19th Nordic Conference of Computational Linguistics (NODALIDA)*, pages 387–474. Linköping University Electronic Press.

Von Klopp, Ana. 1993. *Negation: Implications for theories of natural language*. Ph.D. thesis, University of Edinburgh.

Walker, Marilyn A, Grace I Lin, and Jennifer Sawyer. 2012. An Annotated Corpus of Film Dialogue for Learning and Characterizing Character Style. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1378. European Language Resources Association (ELRA).

A Amplifiers

gradation amplifiers	affirmation amplifiers	surprise amplifiers
very	absolutely	amazingly
awfully	definitely	dramatically
completely	famously	drastically
enormously	genuinely	emphatically
entirely	immaculately	exceptionally
exceedingly	overly	extraordinarily
excessively	perfectly	fantastically
extremely	really	horribly
fucking	severely	incredibly
fuckin	surely	insanely
greatly	thoroughly	phenomenally
highly	truly	remarkably
hugely	undoubtedly	ridiculously
immensely		strikingly
intensely		surprisingly
particularly		terribly
radically		unusually
significantly		wildly
strongly		wonderfully
substantially		amazing
totally		dramatic
utterly		drastic
vastly		emphatic
		exceptional
		extraordinary
		fantastic
		incredible
		phenomenal
		remarkable
		striking
		surprising
		unusual

B Negations

Negations are taken from Quirk et al. (1985, §10.54ff) and Biber (1995).

analytic	no, not
contractions	n't, ain't, aren't, arent, aren't, can't, cannot, cant, couldn't,, didn't, doesn't, don't, hadn't, hasn't, haven't, isn't, mightn't, mustn't, shouldn't, wasn't,, weren't, won't, wouldn't
constructional	neither, never, nor, none, nobody, no-one, without, sans, w/o