11724 HW6: Sociolinguistics

Yingshan Chang
yingshac
yingshac@andrew.cmu.edu

December 6, 2021

1. Introduction

1.1. The Involved-Informational Dimension

Biber1991 [2] introduced five major dimensions of variation in English. This paper studies the first dimension: ***Involved* vs. *Informational***.

The *Involved* style is marked by features which typically show interactions between the writer and the reader. On the other hand, the *Informational* label is usually attributed to formalness. Genres such as professional letters, academic prose and official documents are informational in purpose.

1.2. Hypothesis

Previous works [1, 2] have found consistent differences between male and female authors along this dimension, with female authors strongly tending to the *involved* and male authors to the *informational*. This is because female authors tend to write more conversational, interactive and spontaneous texts. I suspect that these identified characteristics of female authors are also applicable to the young half along the *age* axis. Therefore, I hypothesize that the occurrences of *informational* (*involved*) features linearly increase (decrease) as *age* grows. This paper tests this hypothesis using automated corpus analysis tools and statistical testing.

2. Data

2.1. Blog Authorship Corpus

The Blog Authorship Corpus [1] was gathered from *blogger.com* in August 2004. The corpus incorporates over 680k posts written on or before 2004 and 140M words. All posts are labeled with the blogger's self-provided gender, age, industry (topic), as well as date. A prior work [6] studied this dataset on the relationship between writing style (with respect to lexical features) and gender/age, in light of automated author profiling. This paper explores the relationship between *age* and the *Involved-Informational* variation [2, 4] in English.

[1] The Blog Authorship Corpus is publicly available at https://www.kaggle.com/rtatman/blog-authorship-corpus

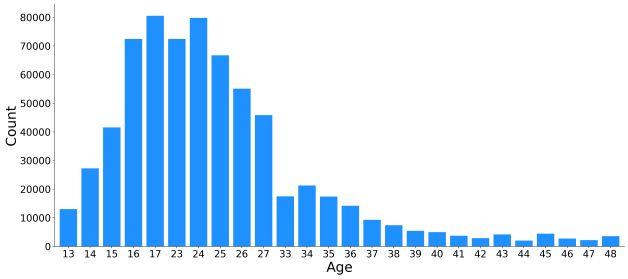

Figure 1. Distribution of #data against the social variable *age*.

2.2. Tagging

POS tags are obtained from Treetagger implemented [2] by Helmut Schmid.

2.3. Data Cleaning

I find that ellipsis ('...') constantly occurs in the dataset. The frequency is great enough to suggest that ellipses were introduced by web-crawling artifacts rather than by the original authors intentionally. Even if ellipses were intentionally written by the original authors as a marker for pause or hesitation, I believe that they should not break the continuation of text when counting bi-gram- or tri-gram-based features. Thus, all ellipses are omitted in the subsequent analysis. Empty data points resulted from ellipsis removal are also ignored. In total, there are 678,103 valid data points.

3. Methodology

3.1. Social Variable

The social variable studied in this paper is *age*. The distribution of age-group sizes is visualized in Fig. 1.

3.2. Linguistic Variables

[3] proposed a list of linguistic features which constitute a good part of the *Involved-Informational* dimension. 12 features are investigated in this paper, whose definitions

[2] A Python module for interfacing with the Treetagger is publicly available at https://github.com/miotto/treetagger-python

and search patterns will be described in detail below. The search patterns are designed based on [5]. Note, NOUNS and ATTRIBUTIVE ADJS correlate with the *informational*, while the rest correlate with the *involved*.

3.2.1 NOUNS

The noun count provides an overall nominal assessment.

Search Pattern Count all nouns, excluding gerunds and nominalizations (i.e. ending with tion|ment|ness|ity).

3.2.2 ATTRIBUTIVE ADJS

An attributive adjective is an adjective not identified as predicative complementing a copular verb (e.g. That's right. The fans became restless.)

Search Pattern Count all adjectives followed by either a noun or another adjective. Considering consecutive adjectives modifying the same noun phrase may be separated by ',' or 'and', the search algorithm is implemented to skip ',' and 'and'.

3.2.3 PRIVATE VERBS

Quirk1985 [4] subdivided factual verbs into "public" and "private" types. The "private" type of factual verbs describe intellectual states such as belief and intellectual acts such as discovery. These states and acts are "private" in the sense that they are not observable.

Search Pattern Quirk1985 [4] provided a list of examples of private verbs. But the list is complete enough to serve as a search template for this feature. All tokens with one of the following lemmas are counted.

> 'accept', 'anticipate', 'ascertain', 'assume', 'believe', 'calculate', 'check', 'conclude', 'conjecture', 'consider', 'decide', 'deduce', 'deem', 'demonstrate', 'determine', 'discern', 'discover', 'doubt', 'dream', 'ensure', 'establish', 'estimate', 'expect', 'fancy', 'fear', 'feel', 'find', 'foresee', 'forget', 'gather', 'guess', 'hear', 'hold', 'hope', 'imagine', 'imply', 'indicate', 'infer', 'insure', 'judge', 'know', 'learn', 'mean', 'note', 'notice', 'observe', 'perceive', 'presume', 'presuppose', 'pretend', 'prove', 'realize', 'reason', 'recall', 'reckon', 'recognize', 'reflect', 'remember', 'reveal', 'see', 'sense', 'show', 'signify', 'suppose', 'suspect', 'think', 'understand'

3.2.4 CONTRACTIONS

There are two major classes of contraction in English: verb contraction and not-contraction. Verb contractions occur with the primary verbs *be* and *have* as well as with the modal verbs *will* and *would* (e.g. I'd, you've, he'll). Not-contractions occur when *not* is reduced and attached to the preceding primary or modal verb (e.g. aren't, didn't, can't).

Search Pattern TreeTagger is able to separate the attachment head from the suffix. In order to exclude the possessive forms of contractions (*'s*), I search for all tokens beginning with an apostrophe and being tagged as either VB* (any verb) or MD (modal verbs).

3.2.5 ANALYTIC NEGATION

Biber1991 [2] distinguished between synthetic negation (e.g. no-negation, neither/nor), which is more literary and integrated, and analytic negation (e.g. not-negation), which is more colloquial and fragmented.

Search Pattern All occurrences of the lemma *not* and *n't* are counted.

3.2.6 PRONOUN 'IT'

According to Quirk 1985 [4], *it* serves both referring and non-referring functions, where the non-referring function is also called "Prop it", used as an empty subject (e.g. What time is it? It's warm today). Biber 1991 [2] treats the use of *it* as a marker for a non-informational focus, due to the fact that *it* can be substituted for nouns, phrases and whole clauses.

Search Pattern All occurrences of the lemma *it* are counted.

3.2.7 CAUSATIVE SUBORDINATION

Biber1991 [2] stated that *because* is the only unambiguous from of causative subordination, whereas other forms such as *as*, *for* and *since* can have a range of functions. Thus, for the ease of automated corpus analysis, I only focus on *because* and leave the investigation of other markers of causative subordination to future work.

Search Pattern All occurrences of the lemma *because* are counted.

3.2.8 PRESENT-TENSE VERBS

Present tense verbs deal with topics and actions of immediate relevance. Besides, cognitive verbs, which describe the writer's mental processes, also typically occur in the present sense. Thus, present tense verbs are markers of the *involved* style.

Search Pattern First, I distinguish the open class of *full verbs* from the closed classes of *primary verbs* (e.g. be, have, do) and *modal verbs* (e.g. will, might). Of these three classes, *full verbs* can act only as main verbs. So all occurrences of VBZ and VBP tags, if the according lemma is not *be*, *have* or *do*, are counted. Note, I exclude verbs in

their base forms, which are tagged as VB. *Modal verbs* are tagged as MD so they can be easily ruled out.

For the remaining cases when the tag is VBZ or VBP:

If the lemma is *be*, count as occurrence if (optionally followed by *not*/*n't*) and not followed by any other verb.

If the lemma is *have*, count as occurrence if (optionally followed by *not*/*n't*) and followed by the verb *get*.

3.2.9 1ST PRONOUNS

First person pronouns are markers of ego-involvement in a text, introducing the writer into the text. First person and second person pronouns together encode the writer-reader relationship specifically into the discourse, indicating an *involved* style.

Search Pattern All occurrences of first person pronouns are counted, omitting the possessive forms *mine* and *ours*.

> 'I', 'me', 'my', 'myself', 'we', 'us', 'our', 'ourselves'

3.2.10 2ND PRONOUNS

Second person pronouns draw the reader into the text. Similar to first person pronouns, second person pronouns indicate a high degree of involvement.

Search Pattern All occurrences of second person pronouns are counted, omitting the possessive form *yours*.

> 'you', 'your', 'yourself', 'yourselves'

3.2.11 INDEFINITE PRONOUNS

According to Quirk1985 [4], indefinite pronouns lack the definiteness which is found in the personal, reflexive, possessive and demonstrative pronouns. Also, indefinite pronouns have universal or partitive meaning, thereby presenting a non-informational focus.

Search Pattern The search algorithm counts all occurrences of the indefinite pronouns Quirk1985 lists.

> 'everyone', 'everything', 'everybody', 'someone', 'somebody', 'something', 'anyone', 'anything', 'anybody', 'nobody', 'nothing', 'none'

3.2.12 AMPLIFIERS

Quirk [4] introduced two subgroups of amplifiers: *maximizers*, denoting the upper extreme of the scale, and *boosters*, denoting a high point on the scale. Biber1991's list of amplifiers corresponds to *maximizers*, plus an additional item *very*, which can premodify maximizers.

Feature (Y)	Linear Model	$P > \|t\|$	R^2
NOUNS	Y = 247.35 + 0.2637X	0.329	0.040
ATTRIBUTIVE ADJS	Y = 28.82 + 0.2109X	0.000	0.561
PRIVATE VERBS	Y = 17.68 - 0.0804X	0.002	0.346
CONTRACTIONS	Y = 15.05 - 0.1393X	0.000	0.425
ANALYTIC NEG	Y = 10.05 - 0.0447X	0.002	0.343
PRONOUN 'IT'	Y = 13.57 - 0.0805X	0.000	0.531
CAUS SUBORD	Y = 1.45 - 0.0125X	0.000	0.521
PRES-TENSE VERBS	Y = 59.37 - 0.3030X	0.000	0.634
1ST PRONOUNS	Y = 53.35 - 0.3266X	0.001	0.374
2ND PRONOUNS	Y = 14.76 - 0.1389X	0.000	0.809
INDEF PRONOUNS	Y = 6.14 - 0.0687X	0.000	0.788
AMPLIFIERS	Y = 1.69 - 0.0056X	0.009	0.253

Table 1. Regression Analysis

Search Pattern The search algorithm counts all occurrences of the amplifiers Biber1991 lists.

> 'absolutely', 'altogether', 'completely', 'enormously', 'entirely', 'extremely', 'fully', 'greatly', 'highly', 'intensely', 'perfectly', 'strongly', 'thoroughly', 'totally', 'utterly', 'very'

3.3. Statistical Tests

The occurrences of linguistic features are counted (according to the search patterns in Sec. 3.2), grouped by *age* and averaged. I perform simple linear regression with *age* being the independent variable and feature count being the dependent variable.

4. Results

The t-test examines whether the coefficient before the independent variable is significant. The significance threshold for this study is set to 0.01.

R^2 is a goodness-of-fit measure for linear regression models. It signifies the percentage of variance in the dependent variable explained by the independent variable. For example, R^2 = 0.346 for Y = PRIVATE VERBS means about 34.6% of the variability of PRIVATE VERBS is explained by *age*. $R^2 > 0.25$ is commonly considered as an indicator that the regression model provides adequate fit to data.

Table. 1 summarizes the regression analyses between each linguistic feature and *age*. The regression coefficient and goodness-of-fit are significant for all linguistic features except for NOUNS. Also, the sign of each regression coefficient is consistent with whether the feature is identified as *involved* or *informational*. Data and regression lines are visualized in Fig. 2.

5. Discussion

11 out of the 12 regression analyses are significant, strongly supporting the hypothesis that elder peo-

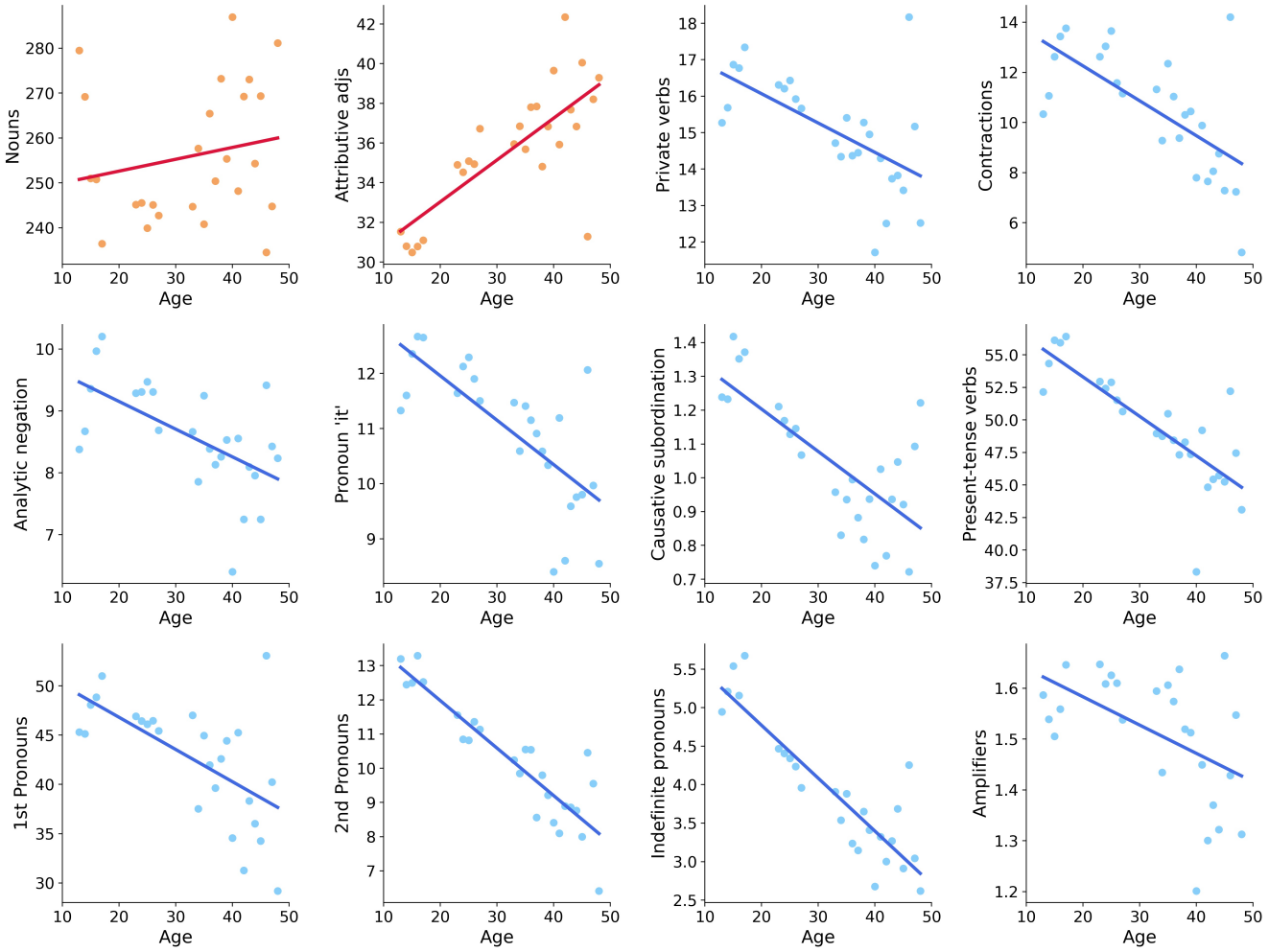

Figure 2. Plot the mean feature count per 1000 tokens against *age* for 12 linguistic features. The linear regression models for all linguistic features except for NOUNS are significant.

ple write more informational and less involved English than younger people do. ATTRIBUTIVE ADJS, PRONOUN 'IT', CAUSATIVE SUBORDINATION and PRESENT-TENSE VERBS have moderate correlations (>0.5) with *age*. 2ND PRONOUNS and INDEFINITE PRONOUNS have high correlations (>0.75) with *age*.

Limitations This Blog Authorship Corpus contains a lot of emoticons such as ":-)", "=)", which will be separated by TreeTagger into independent punctuations. It is difficult to systematically remove emojis. For the ease of experimentation I did not clean up emoticons in the pre-processing stage. So, there might be noise creeping into the subsequent analysis. It is also worth noting that conclusions drawn from this paper is more applicable to informal written English and should be taken with a grain of salt under other domains. Another limitation of this study lies in the unbalanced group sizes. Small groups in the tail are more vulnerable to outliers. According to Fig. 2, dots on the elder half of the age axis tend to deviate more from the corresponding regression line, which might be due to less faithful estimation of population statistics from a small sample.

Future Directions The Blog Authorship Corpus also provides a good venue for investigating the usage of slangs and neologism by people in different *age* groups. Example features are as follows.

Emoticons e.g. ":-)", "=)".
Swear words e.g. "Damnit.
Abbreviations e.g. "b/c", "b/4", "Very Happy B'day".
Interjections e.g. "woah", "*meh*", "Heehee", "YaY!!!".
Capitalization e.g. "DONT THINK IM KIDDING YOU", "I HATE BAND! END OF STORY!".

References

[1] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text & Talk*, 23(3):321–346, 2003. 1

[2] Douglas Biber. *Variation across speech and writing*. Cambridge University Press, 1991. 1, 2

[3] Douglas Douglas. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, 26(5):331–345, 1992. 1

[4] Randolph Quirk, Sidney Greenbaum, and Geoffrey Leech. J. svartvik (1985) a comprehensive grammar of the english language. *Harlow: Longman*, 8. 1, 2, 3

[5] Marc Reymann. *The textual dimension" Involved-Informational": A corpus-based study*. PhD thesis, 2005. 2

[6] J Schler, M Koppel, S Argamon, and JW Pennebaker. Effects of age and gender on blogging in proceedings of 2006 aaai spring symposium on computational approaches for analyzing weblogs. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 2006. 1