Cross-lingual syntactic variation over age and gender

Anders Johannsen, Dirk Hovy, and Anders Søgaard
Center for Language Technology
University of Copenhagen, Denmark
Njalsgade 140
{ajohannsen,dirk.hovy,soegaard}@hum.ku.dk

Abstract

Most computational sociolinguistics studies have focused on *phonological* and *lexical* variation. We present the first large-scale study of *syntactic* variation among demographic groups (age and gender) across several languages. We harvest data from online user-review sites and parse it with universal dependencies. We show that several age and gender-specific variations hold across languages, for example that women are more likely to use VP conjunctions.

1 Introduction

Language varies between demographic groups. To detect this variation, sociolinguistic studies require *both* a representative corpus of text *and* meta-information about the speakers. Traditionally, this data was collected from a combination of interview transcriptions and questionnaires. Both methods are time-consuming, so population sizes have been small, sometimes including less than five subjects (Rickford and Price, 2013). While these resources enable detailed qualitative analyses, small sample sizes may lead to false research findings (Button et al., 2013). Sociolinguistic studies, in other words, often lack statistical power to establish relationships between language use and socio-economic variables.

Obtaining large enough data sets becomes even more challenging the more complex the target variables are. So while syntactic variation has been identified as an important factor of variation (Cheshire, 2005), it was not approached, due to its high complexity. This paper addresses the issue systematically on a large scale. In contrast to previous work in both sociolinguistics and NLP, we consider syntactic variation across groups at the level of *treelets*, as defined by dependency structures, and make use of a large corpus that includes demographic information on both age and gender.

The impact of such findings goes beyond sociolinguistic insights: knowledge about systematic differences among demographic groups can help us build better and fairer NLP tools. Volkova et al. (2013), Hovy and Søgaard (2015), Jørgensen et al. (2015), and Hovy (2015) have shown the impact of demographic factors on NLP performance. Recently, the company Textio introduced a tool to help phrase job advertisements in a gender-neutral way.[1] While their tool addresses lexical variation, our results indicate that linguistic differences extend to the syntactic level.

Previous work on demographic variation in both sociolinguistics and NLP has begun to rely on corpora from social media, most prominently Twitter. Twitter offers a sufficiently large data source with broad coverage (albeit limited to users with access to social media). Indeed, results show that this resource reflects the *phonological* and *morpholexical* variation of spoken language (Eisenstein, 2013b; Eisenstein, 2013a; Doyle, 2014).

However, Twitter is not well-suited for the study of *syntactic* variation for two reasons. First, the limited length of the posts compels the users to adopt a terse style that leaves out many grammatical markers. As a consequence, performance of syntactic parsers is prohibitive for linguistic analysis in this domain. Second, Twitter provides little meta-information about the users, except for regional origin and time of posting. Existing work has thus been restricted to these demographic variables. One line of research has focused on predictive models for age and gender (Alowibdi et al., 2013; Ciot et al., 2013) to add meta-data on Twitter, but again, error rates are too high for use in sociolinguistic hypothesis testing.

We use a new source of data, namely the user

[1] http://recode.net/2015/04/20/textio-spell-checks-for-gender-bias/

Proceedings of the 19th Conference on Computational Language Learning, pages 103–112,
Beijing, China, July 30-31, 2015.

review site Trustpilot. The meta-information on Trustpilot is both more prevalent and more reliable, and textual data is not restricted in length (see Table 2). We use state-of-the-art dependency parsers trained on universal treebanks (McDonald et al., 2013) to obtain comparable syntactic analyses across several different languages and demographics.

Contributions We present the first study of morpho-syntactic variation with respect to demographic variables across several languages at a large scale. We collect syntactic features within demographic groups and analyze them to retrieve the most significant differences. For the analysis we use a method that preserves statistical power, even when the number of possible syntactic features is very large. Our results show that demographic differences extend *beyond* lexical choice.

2 Data collection

The TRUSTPILOT CORPUS consists of user reviews from the Trustpilot website. On Trustpilot, users can review company websites and leave a one to five star rating, as well as a written review. The data is available for 24 countries, using 13 different languages (Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish). In our study, we are limited by the availability of comparable syntactically annotated corpora (McDonald et al., 2013) for five languages used in eleven countries, i.e., English (Australia, Canada, UK, and US), French (Belgium and France), German (Switzerland and Germany), Italian, Spanish, and Swedish. We treat the different variants of these languages separately in the experiments below.[2]

Many users opt to provide a public profile. There are no mandatory fields, other than name, but many also supply their birth year, gender, and location. We crawl the publicly available information on the web site for users and reviews, with different fields. Table 1 contains a list of the fields that are available for each type of entity. For more information on the data as a source for demographic information, see Hovy et al. (2015).

We enhance the data set for our analysis by adding gender information based on first names. In order to add missing gender information, we measure the distribution over genders for each name. If a name occurs with sufficient frequency and is found predominantly in one gender, we propagate this gender to all occurrences of the name that lack gender information. In our experiments, we used a gender-purity factor of 0.95 (name occurs with one gender 95% of the time) and a minimum frequency of 3 (name appears at least 3 times in the data). Since names are language-specific (*Angel* is male in Spanish, but female in English), we run this step separately on each language. On average, this measure doubled the amount of gender information for a language.

Note that the domain (reviews) potentially introduces a bias, but since our analysis is largely at the syntactic level, we expect the effect to be limited. While there is certainly a domain effect at the lexical level, we assume that the syntactic findings generalize better to other domains.

Users	Name, ID, profile text, location (city and country), gender, year of birth
Reviews	Title, text, rating (1–5), User ID, Company ID, Date and time of review

Table 1: Meta-information in TRUSTPILOT data

	Users	Age	Gender	Place	All
UK	1,424k	7%	62%	5%	4%
France	741k	3%	53%	2%	1%
Denmark	671k	23%	87%	17%	16%
US	648k	8%	59%	7%	4%
Netherlands	592k	9%	39%	7%	5%
Germany	329k	8%	47%	6%	4%
Sweden	170k	5%	64%	4%	3%
Italy	132k	10%	61%	8%	6%
Spain	56k	6%	37%	5%	3%
Norway	51k	5%	50%	4%	3%
Belgium	36k	13%	42%	11%	8%
Australia	31k	8%	36%	7%	5%
Finland	16k	6%	36%	5%	3%
Austria	15k	10%	43%	7%	5%
Switzerland	14k	8%	41%	7%	4%
Canada	12k	10%	19%	9%	4%
Ireland	12k	8%	30%	7%	4%

Table 2: No. of users per variable per country (after augmentations), for countries with $10k+$ users.

3 Methodology

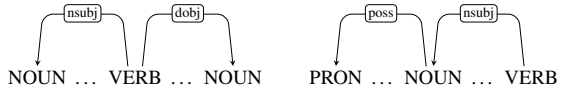
For each language, we train a state-of-the-art dependency parser (Martins et al., 2013) on a

[2]While this might miss some dialectal idiosyncrasies, it is based on standard NLP practice, e.g., when using WSJ-trained parsers in translation of (British) Europarl.

treebank annotated with the Stanford dependency labels (McDonald et al., 2013) and universal POS tag set (Petrov et al., 2011). This gives us syntactic analyses across all languages that describe the same syntactic phenomena the same way. Figure 1 shows two corpus sentences annotated with this harmonized representation.

The style of the reviews is much more canonical than social web data, say Twitter. Expected parse performance can be estimated from the SANCL 2012 shared task on dependency parsing of web data (Petrov and McDonald, 2012). The best result on the review domain there was 83.86 LAS and 88.31 UAS, close to the average over all web domains (83.45 LAS and 87.62 UAS).

From the parses, we extract all subtrees of up to three tokens (*treelets*). We do not distinguish between right- and left-branching relations: the representation is basically a "bag of relations". The purpose of this is to increase comparability across languages with different word orderings (Naseem et al., 2012). A one-token treelet is simply the POS tag of the token, e.g. NOUN or VERB. A two-token treelet is a typed relation between head and dependent, e.g. VERB $\xrightarrow{\text{NSUBJ}}$ NOUN. Treelets of three tokens have two possible structures. Either the head directly dominates two tokens, or the tokens are linked together in a chain, as shown below:



3.1 Treelet reduction

We extract between 500,000 to a million distinct treelets for each language. In principle, we could directly check for significant differences in the demographic groups and use Bonferroni correction to control the family-wise error (i.e., the probability of obtaining a false positive). However, given the large number of treelets, the correction for multiple comparisons would underpower our analyses and potentially cause us to miss many significant differences. We therefore reduce the number of treelets by two methods.

First, we set the minimum number of occurrences of a feature in each language to 50. We apply this heuristic both to ensure statistical power and to focus our analyses on prevalent rather than rare syntactic phenomena.

Second, we perform feature selection using L_1 randomized logistic regression models, with age or gender as target variable, and the treelets as input features. However, direct feature selection with L_1 regularized models (Ng, 2004) is problematic when variables are highly correlated (as in our treelets, where e.g. three-token structures can subsume smaller ones). As a result, small and inessential variations in the dataset can determine which of the variables are selected to represent the group, so we end up with random within-group feature selection.

We therefore use *stability selection* (Meinshausen and Bühlmann, 2010). Stability selection mitigates the correlation problem by fitting the logistic regression model hundreds of times with perturbed data (75% subsampling and feature-wise regularization scaling). Features that receive non-zero weights across many runs can be assumed to be highly indicative. Stability selection thus gives *all* features a chance to be selected. It controls the false positive rate, which is less conservative than family-wise error. We use the default parameters of a publicly available stability selection implementation[3], run it on the whole data set, and discard features selected less than 50% of the time.

With the reduced feature set, we check for usage differences in demographic groups (age and gender) using a χ^2 test. We distinguish two age groups: speakers that are younger than 35, and speakers older than 45. These thresholds were chosen to balance the size of both groups. At this stage we set the desired p-value at 0.02 and apply Bonferroni correction, effectively dividing the p-value threshold by the number of remaining treelets.[4]

Note, finally, that the average number of words written by a reviewer differs between the demographic groups (younger users tend to write more than older ones, women more than men). To counteract this effect, the expected counts in our null hypothesis use the proportion of *words* written by people in a group, rather than the proportion of *people* in the group (which would skew the results towards the groups with longer reviews).

[3] `http://scikit-learn.org/`

[4] Choosing a p-value is somewhat arbitrary. Effectively, our p-value cutoff is several orders of magnitude lower than 0.02, due to the Bonferroni correction.

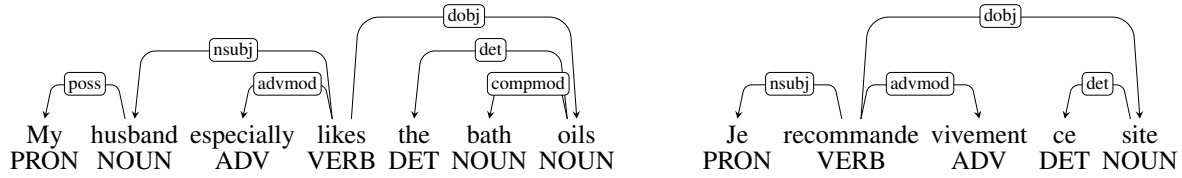


Figure 1: Universal dependency relations for an English and a French sentence, both with adverbial modifiers.

Signif. in # lang.	Rank	Feature	Effect High	Effect By	Subsumes
11	1	NUM	M	32 %	
	2	PRON	F	11 %	
	3	NOUN	M	6 %	
10	4	VERB $\xrightarrow{\text{ACOMP}}$ ADJ	F	22 %	5
	5	VERB	F	6 %	
9	6	ADJ $\xleftarrow{\text{ACOMP}}$ VERB $\xrightarrow{\text{CONJ}}$ VERB	F	36 %	4, 5, 14
	7	VERB $\xrightarrow{\text{ACOMP}}$ ADJ $\xrightarrow{\text{ADVMOD}}$ ADV	F	35 %	4, 5
	8	NOUN $\xrightarrow{\text{COMPMOD}}$ NOUN	M	22 %	3
	9	VERB $\xrightarrow{\text{NSUBJ}}$ PRON	F	14 %	2, 5
8	10	VERB $\xrightarrow{\text{CONJ}}$ VERB $\xrightarrow{\text{ACOMP}}$ ADJ	F	40 %	4, 5, 14
	11	VERB $\xrightarrow{\text{ACOMP}}$ ADJ $\xrightarrow{\text{CONJ}}$ ADJ	F	36 %	4, 5
	12	ADJ $\xleftarrow{\text{ACOMP}}$ VERB $\xrightarrow{\text{CC}}$ CONJ	F	28 %	4, 5
	13	CONJ $\xleftarrow{\text{CC}}$ VERB $\xrightarrow{\text{CONJ}}$ VERB	F	16 %	5, 14
	14	VERB $\xrightarrow{\text{CONJ}}$ VERB	F	14 %	5
	15	ADP $\xleftarrow{\text{ADPMOD}}$ VERB $\xrightarrow{\text{NSUBJ}}$ NOUN	M	14 %	3, 5
	16	NOUN $\xrightarrow{\text{ADPMOD}}$ ADP $\xrightarrow{\text{ADPOBJ}}$ NOUN	M	13 %	3, 17
	17	NOUN $\xrightarrow{\text{ADPMOD}}$ ADP	M	13 %	3
	18	VERB $\xrightarrow{\text{AUX}}$ VERB	F	10 %	5
7	19	ADP $\xrightarrow{\text{ADPOBJ}}$ NUM	M	43 %	1
	20	ADJ $\xleftarrow{\text{ACOMP}}$ VERB $\xrightarrow{\text{NSUBJ}}$ PRON	F	41 %	2, 4, 5, 9

Table 3: **Gender comparison**: Significant syntactic features across languages. Features ordered by number of languages in which they are significant. Right-hand side shows the gender for which the feature is indicative, by which margin, and whether it subsumes other features (indexed by rank)

4 Results

We are interested in robust syntactic variation across languages; that is, patterns that hold across most or all of the languages considered here. We therefore score each of the identified treelets by the number of languages with a significant difference in occurrence between the groups of the given demographic variable. Again, we use a rather conservative non-parametric hypothesis test, with Bonferroni correction.

Tables 3 and 4 show the results for age and gender, respectively. The first column shows the number of languages in which the treelet (third column) is significant. The fourth and fifth column indicate for which age or gender subgroup the feature is indicative, and how much larger the rate of occurrence is there in percent. The indices in the last column represent containment relationships, i.e., when a treelet is strictly contained in another treelet (indexed by the rank given in the second column).

In the case of gender, three atomic treelets (parts of speech) correlate significantly across *all* 11 languages. Two treelets correlate significantly across 10 languages. For age, five treelets correlate significantly across 10 languages.

In sum, men seem to use numerals and nouns more than women across languages, whereas women use pronouns and verbs more often. Men use nominal compounds more often than women in nine out of eleven languages. Women, on the other hand, use VP coordinations more in eight out of eleven languages.

For age, some of the more striking patterns

involve prepositional phrases, which see higher use in the older age group. In atomic treelets, noun use is slightly higher in the older group, while pronouns are more often used by younger reviewers.

Our results address a central question in variational linguistics, namely whether syntax plays a role in language variation among groups. While this has been long suspected, it was never empirically researched due to the perceived complexity. Our findings are the first to corroborate the hypotheses that language variation goes beyond the lexical level.

	FR	DE	IT	ES	SE	UK	US
FR	**592**	42%	73%	88%	46%	47%	38%
DE		**365**	46%	56%	45%	52%	43%
IT			**138**	50%	32%	72%	65%
ES				**78**	28%	79%	73%
SE					**182**	49%	45%
UK						**1056**	88%
US							**630**

	FR	DE	UK	US
FR	**108**	57%	53%	35%
DE		**237**	46%	27%
UK			**370**	56%
US				**173**

Table 5: **Gender (top) and age (bottom)**: Pairwise overlap in significant features. Languages with 50 or less significant features were left out. Diagonal gives the number of features per language.

We also present the pairwise overlap in significant treelets between (a subset of the) languages. See Table 5. Their diagonal values give the number of significant treelets for that language. Percentages in the pairwise comparisons are normalized by the smallest of the pair. For instance, the 49 % overlap between Sweden (SE) and United Kingdom (UK) in Table 5 means that 49 % of the 182 SE treelets were also significant in UK.

We observe that English variants (UK and US) share many features. The Romance languages also share many features with each other, but Italian and Spanish also share many features with English. In Section 5, we analyze our results in more depth.

5 Analysis of syntactic variation

Due to space constraints, we restrict our analysis to a few select treelets with good coverage and interpretable results.

5.1 Gender differences

The top features for gender differences are mostly atomic (pre-terminals), indicating that we observe the same effect as mentioned previously in the literature (Schler et al., 2006), namely that certain parts-of-speech are prevalent in one gender.

1, 2, 3 For all languages, the use of numerals and nouns is significantly correlated with men, while pronouns and verbs are more indicative of women. When looking at the types of pronouns used by men and women, we see very similar distributions, but men tend to use impersonal pronouns (*it*, *what*) more than women do. Nouns and numbers are associated with the alleged "information emphasis" of male language use (Schler et al., 2006). Numbers typically indicate prices or model numbers, while nouns are usually company names.

The robustness of POS features could to some extent be explained by the different company categories reviewed by each gender: in COMPUTER & ACCESSORIES and CAR LIGHTS the reviews are predominately by men, while the reviews in the PETS and CLOTHES & FASHION categories are mainly posted by women. Using numerals and nouns is more likely when talking about computers and car lights than when talking about pets and clothing, for example.

4 In English, this treelet is instantiated by examples such as:

(1) is/was/are great/quick/easy and is/was/arrived

In German, the corresponding examples would be:

(2) bin/war zufrieden und werde/würde wieder bestellen (am/was satisfied and will/would order again)

8 This feature mainly encompasses noun compounds, incl., company names. Again, this feature is indicative of male language use. This may be a side-effect of male use of nouns, but note that the effect is much larger with noun compounds.

Signif. in # lang.	Rank	Feature	Effect High	Effect By	Subsumes
8	[1]	NOUN	>45	5 %	
7	[2]	ADP $\xrightarrow{\text{ADPOBJ}}$ NOUN $\xrightarrow{\text{ADPMOD}}$ ADP	>45	20 %	[1], [5], [9]
	[3]	NOUN $\xrightarrow{\text{ADPMOD}}$ ADP $\xrightarrow{\text{ADPOBJ}}$ NOUN	>45	14 %	[1], [5], [9]
	[4]	VERB $\xrightarrow{\text{ADVMOD}}$ ADV	<35	12 %	
	[5]	ADP $\xrightarrow{\text{ADPOBJ}}$ NOUN	>45	8 %	[1]
6	[6]	ADV $\xleftarrow{\text{ADVMOD}}$ VERB $\xrightarrow{\text{CONJ}}$ VERB	<35	34 %	[4], [19]
	[7]	VERB $\xleftarrow{\text{ADVCL}}$ VERB $\xrightarrow{\text{ADVMOD}}$ ADV	<35	27 %	[4], [20]
	[8]	VERB $\xrightarrow{\text{CC}}$ CONJ	<35	15 %	
	[9]	NOUN $\xrightarrow{\text{ADPMOD}}$ ADP	>45	12 %	[1]
	[10]	PRON	<35	10 %	
5	[11]	ADP $\xleftarrow{\text{ADPMOD}}$ NOUN $\xrightarrow{\text{COMPMOD}}$ NOUN	>45	40 %	[1], [9], [18]
	[12]	VERB $\xrightarrow{\text{CONJ}}$ VERB $\xrightarrow{\text{NSUBJ}}$ PRON	<35	32 %	[10], [19]
	[13]	ADV $\xleftarrow{\text{ADVMOD}}$ VERB $\xrightarrow{\text{CC}}$ CONJ	<35	25 %	[4], [8]
	[14]	ADP $\xrightarrow{\text{ADPOBJ}}$ NOUN $\xrightarrow{\text{COMPMOD}}$ NOUN	>45	23 %	[1], [5], [18]
	[15]	CONJ $\xleftarrow{\text{CC}}$ VERB $\xrightarrow{\text{NSUBJ}}$ PRON	<35	21 %	[8], [10]
	[16]	CONJ $\xleftarrow{\text{CC}}$ VERB $\xrightarrow{\text{CONJ}}$ VERB	<35	20 %	[8], [19]
	[17]	ADV $\xleftarrow{\text{ADVMOD}}$ VERB $\xrightarrow{\text{NSUBJ}}$ PRON	<35	19 %	[4], [10]
	[18]	NOUN $\xrightarrow{\text{COMPMOD}}$ NOUN	>45	17 %	[1]
	[19]	VERB $\xrightarrow{\text{CONJ}}$ VERB	<35	16 %	
	[20]	VERB $\xrightarrow{\text{ADVCL}}$ VERB	<35	11 %	

Table 4: **Age group comparison**: Significant syntactic features across languages. Layout as in Table 3

5.2 Age differences

For age, features vary a lot more than for gender, i.e., there is less support for each than there was for the gender features. A few patterns still stand out.

[2] This pattern, which is mostly used by the > 45 age group, is often realized in English to express temporal relations, such as

(1) (with)in a couple/days/hours of
(2) in time for

In German, it is mostly used to express comparisons

(1) im Vergleich/Gegensatz zu (compared/in contrast to)
(2) auf Suche nach (in search of)
(3) in Höhe/im Wert von (valued at)

[3] This pattern, which is indicative of the > 45 age group, is mostly realized in English to express a range of prepositional phrases, some of them overlapping with the previous pattern:

(1) value for money
(2) couple of days
(3) range of products

German also shows prepositional phrases, yet no overlap with [2]

(1) Qualität zu Preisen (quality for price)
(2) Auswahl an Weinen/Hotels (selection of wines/hotels)

In French, this mostly talks about delivery

(1) délai(s) de livraison (delivery)
(2) rapidité de livraison (speed of delivery)

And in Spanish, the main contenders are complex (and slightly more formal) expressions

(1) gastos de envío (shipping)
(2) atención al cliente (customer service)

[4] This pattern is mostly used by the younger group, and realized to express positive recommendations in all languages:

(1) use again/definitely
(2) recommend highly/definitely

German:

(1) empfehle nur/sehr (just recommend)
(2) bestelle wieder/dort/schon (order again/there/already)

French:

(1) recommande vivement (vividly recommend)
(2) emballé/passé/fait bien (packaged/delivered/made well)

[5] This pattern is again predominant in the older group, and mostly used in English to complement the prepositional phrases in [3]

(1) at price
(2) with service

In German, it is mostly used to express comparisons

(1) in Ordnung (alright)
(2) am Tag (on the day)

6 Semantic variation within syntactic categories

Given that a number of the indicative features are single treelets (POS tags), we wondered whether there are certain semantic categories that fill these slots. Since we work across several languages, we are looking for semantically equivalent classes. We collect the most significant adjectives and adverbs for each gender for each language and map the words to all of their possible lexical groups in BabelNet (Navigli and Ponzetto, 2010). This creates lexical equivalence classes. Table 6 shows the results. We purposefully exclude nouns and verbs here, as there is too much variation to detect any patterns.

The number of languages that share lexical items from the same BabelNet class is typically smaller than the number of languages that share a treelet. Nevertheless, we observe certain patterns.

The results for gender are presented in Table 6. For adverbs, the division seems to be about intensity: men use more *downtoners* (*approximately*; *almost*; *still*), while women use more *intensifiers* (*actually*; *really*; *truly*; *quite*; *lots*). This finding is new, in that it directly contradicts the perceived wisdom of female language as being more restrained and hedging.

In their use of adjectives, on the other hand, men highlight "factual" properties of the subject, such as price (*inexpensive*) and quality (*cheap*; *best*; *professional*), whereas women use more qualitative adjectives that express the speaker's opinion about the subject (*fantastic*; *amazing*; *pretty*) or their own state (*happy*), although we also find the "factual" assessment *simple*.

Table 7 shows the results for age. There are not many adjectives that group together, and they do not show a clear pattern. Most of the adverbs are indicative of the younger group, although there is overlap with the older group (this is due to different sets of words mapping to the same class). We did not find any evidence for pervasive age effects across languages.

Langs.	BABELNET class	Highest
	Adverbs	
5	just about; approximately	M
	actually; indeed	F
	real; really; very	F
	really; truly; genuinely	F
	quite	F
4	almost; nearly; virtually	M
	still	M
	however; still; nevertheless	M
	soon; presently; shortly	F
	a good deal; lots; very much	F
	Adjectives	
6	fantastic; wondrous; wonderful	F
5	inexpensive; cheap; economic	M
	amazing; awesome; marvelous	F
	tinny; bum; cheap	M
4	happy	F
	best (quality)	M
	professional	M
	pretty	F
	easy; convenient; simple	F
	okay; o.k.; all right	M

Table 6: **Gender**: equivalence classes in BabelNet

7 Related Work

Sociolinguistic studies investigate the relation between a speaker's linguistic choices and socio-economic variables. This includes regional origin (Schmidt and Herrgen, 2001; Nerbonne, 2003; Wieling et al., 2011), age (Barke, 2000; Barbieri, 2008; Rickford and Price, 2013), gender (Holmes, 1997; Rickford and Price, 2013), social class (Labov, 1964; Milroy and Milroy, 1992; Macaulay, 2001; Macaulay, 2002), and ethnicity (Carter, 2013; Rickford and Price, 2013). We focus on age and gender in this work.

Corpus-based studies of variation have largely been conducted either by testing for the presence or absence of a set of pre-defined words (Pennebaker et al., 2001; Pennebaker et al., 2003), or by analysis of the unigram distribution (Barbieri, 2008). This approach restricts the findings to the phenomena defined in the hypothesis, in this case the word list used. In contrast, our approach works beyond the lexical level, is data-driven and thus unconstrained by prior hypotheses.

Eisenstein et al. (2011) use multi-output

Langs.	BABELNET class	Highest
	Adverbs	
5	actually; really; in fact	<35
	truly; genuinely; really	<35
4	however; nevertheless	<35
	however	<35
3	merely; simply; just	<35
	reasonably; moderately; fairly	<35
	very	>45
	in truth; really	<35
	very; really; real	>45
	very; really; real	<35
	Adjectives	
3	easy; convenient; simple	<35
	quick; speedy	>45
	costly; pricy; expensive	<35
	simple	<35
	excellent; first-class	>45
2	spacious; wide	<35
	expensive	<35
	simple (unornamented)	<35
	new	>45
	best	<35

Table 7: **Age**: Lexical equivalences in BabelNet

regression to predict demographic attributes from term frequencies, and vice versa. Using sparsity-inducing priors, they identify key lexical variations between linguistic communities. While they mention syntactic variation as possible future work, their method has not yet been applied to syntactically parsed data. Our method is simpler than theirs, yet goes beyond words. We learn demographic attributes from raw counts of syntactic treelets rather than term frequencies, and test for group differences between the most predictive treelets and the demographic variables. We also use a sparsity-inducing regularizer.

Kendall et al. (2011) study dative alternations on a 250k-words corpus of transcribed spoken Afro-American Vernacular English. They use logistic regression to correlate syntactic features and dialect, similar to Eisenstein et al. (2011), but their study differs from ours in using manually annotated data, studying only one dialect and demographic variable, and using much less data.

Stewart (2014) uses POS tags to study morpho-syntactic features of Afro-American Vernacular English on Twitter, such as copula deletion, habitual *be*, null genitive marking, etc. Our study is different from his in using full syntactic analyses, studying variation across age and gender rather than ethnicity, and in studying syntactic variation across several languages.

8 Conclusion

Syntax has been identified as an important factor in language variation among groups, but not addressed. Previous work has been limited by data size or availability of demographic meta-data. Existing studies on variation have thus mostly focused on lexical and phonological variation.

In contrast, we study the effect of age and gender on syntactic variation across several languages. We use a large-scale data source (international user-review websites) and parse the data, using the same formalisms to maximize comparability. We find several highly significant age- and gender-specific syntactic patterns.

As NLP applications for social media become more widespread, we need to address their performance issues. Our findings suggest that including extra-linguistic factors (which become more and more available) could help improve performance of these systems. This requires a discussion of approaches to corpora construction and the development of new models.

Acknowledgements

We would like to thank the anonymous reviewers for their comments that helped improve the paper. This research is funded by the ERC Starting Grant LOWLANDS No. 313695.

References

Jalal S Alowibdi, Ugo A Buy, and Philip Yu. 2013. Empirical evaluation of profile characteristics for gender classification on twitter. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 1, pages 365–369. IEEE.

Federica Barbieri. 2008. Patterns of age-based linguistic variation in American English. *Journal of sociolinguistics*, 12(1):58–88.

Andrew J Barke. 2000. The Effect of Age on the Style of Discourse among Japanese Women. In *Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation*, pages 23–34.

Katherine Button, John Ioannidis, Claire Mokrysz, Brian Nosek, Jonathan Flint, Emma Robinson, and

Marcus Munafo. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14:365–376.

Phillip M Carter. 2013. Shared spaces, shared structures: Latino social formation and african american english in the us south. *Journal of Sociolinguistics*, 17(1):66–92.

Jenny Cheshire. 2005. Syntactic variation and beyond: Gender and social class variation in the use of discourse-new markers1. *Journal of Sociolinguistics*, 9(4):479–508.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of twitter users in non-english contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash*, pages 18–21.

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In *EACL*.

Jacob Eisenstein, Noah Smith, and Eric Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of ACL*.

Jacob Eisenstein. 2013a. Phonological factors in social media writing. In *Workshop on Language Analysis in Social Media, NAACL*.

Jacob Eisenstein. 2013b. What to do about bad language on the internet. In *Proceedings of NAACL*.

Janet Holmes. 1997. Women, language and identity. *Journal of Sociolinguistics*, 1(2):195–223.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of ACL*.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review-sites as a source for large-scale sociolinguistic studies. In *Proceedings of WWW*.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of ACL*.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Workshop on Noisy User-generated Text (W-NUT)*.

Tyler Kendall, Joan Bresnan, and Gerard van Herk. 2011. The dative alternation in african american english. *Corpus Linguistics and Linguistic Theory*, 7(2):229–244.

William Labov. 1964. *The social stratification of English in New York City.* Ph.D. thesis, Columbia university.

Ronald Macaulay. 2001. You're like 'why not?' the quotative expressions of glasgow adolescents. *Journal of Sociolinguistics*, 5(1):3–21.

Ronald Macaulay. 2002. Extremely interesting, very interesting, or only quite interesting? adverbs and social class. *Journal of Sociolinguistics*, 6(3):398–417.

André FT Martins, Miguel Almeida, and Noah A Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *ACL (2)*, pages 617–622.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *ACL*.

Nicolai Meinshausen and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Lesley Milroy and James Milroy. 1992. Social network and social class: Toward an integrated sociolinguistic model. *Language in society*, 21(01):1–26.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 629–637. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

John Nerbonne. 2003. Linguistic variation and computation. In *Proceedings of EACL*, pages 3–10. Association for Computational Linguistics.

Andrew Y Ng. 2004. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, volume 59. Citeseer.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. CoRR abs/1104.2086.

John Rickford and Mackenzie Price. 2013. Girlz ii women: Age-grading, language change and stylistic variation. *Journal of Sociolinguistics*, 17(2):143–179.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205.

Jürgen Erich Schmidt and Joachim Herrgen. 2001. Digitaler Wenker-Atlas (DiWA). Bearbeitet von Alfred Lameli, Tanja Giessler, Roland Kehrein, Alexandra Lenz, Karl-Heinz Müller, Jost Nickel, Christoph Purschke und Stefan Rabanus. Erste vollständige Ausgabe von Georg Wenkers "Sprachatlas des Deutschen Reichs".

Ian Stewart. 2014. Now we stronger than ever: African-american syntax in twitter. *EACL 2014*, page 31.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of EMNLP*, pages 1815–1827.

Martijn Wieling, John Nerbonne, and R Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS one*, 6(9):e23613.