Stylistic variation over 200 years of court proceedings according to gender and social class

Stefania Degaetano-Ortlieb
Department of Language Science and Technology
Saarland University
66123 Saarbrücken, Germany
s.degaetano@mx.uni-saarland.de

Abstract

We present an approach to detect stylistic variation across social variables (here: gender and social class), considering also diachronic change in language use. For detection of stylistic variation, we use relative entropy, measuring the difference between probability distributions at different linguistic levels (here: lexis and grammar). In addition, by relative entropy, we can determine which linguistic units are related to stylistic variation.

1 Introduction

Understanding language/stylistic variation[1] according to social variables (such as gender, age or social class) is of great interest to sociolinguistics (Eckert, 1989; Labov, 1963; Bernstein, 1971; Tagliamonte, 2006) and has recently received increased attention in the NLP community for developing methods able to predict social context based on language use (see Nguyen et al. (2016) for an overview).

In this paper, we take a diachronic perspective and study how language use in court proceedings changes over a time span of approx. 200 years considering the interaction between gender and social class. A major focus is on female of higher class, as we hypothesize that as the inferior social position of women was increasingly questioned from the mid-nineteenth century, this might be reflected in their language use[2]. For this, we use the Old Bailey Corpus (Huber et al., 2016), a diachronic corpus of manually socio-linguistically annotated data of court proceedings raging from 1720 to 1913 (see Section 3.1). We apply an information-theoretic approach using relative entropy, which has been successfully applied for the analysis of diachronic variation in language use investigating the development of written scientific English (cf. Degaetano-Ortlieb et al. (to appear); Degaetano-Ortlieb and Teich (2016)).

We make two major contributions: First, we investigate change in language use showing how groups of gender and of lower vs. higher social class change linguistically over time. This contributes not only to (historical) sociolinguistics but also to the NLP community strengthening awareness of accounting for stylistic variation and diachronic change in language use. Second, rather than selecting predefined features to analyze stylistic variation, we use whole linguistic levels (lexis and grammar) as described in Section 3.2 from which stylistic features can be inferred.

After introducing related work (Section 2) as well as our data set and methodology (Section 3.2), we test our hypothesis of change in language use for female of higher class investigating stylistic variation (Section 4). Section 5 concludes the paper with a brief summary and an outlook on future work.

2 Related Work

Traditional sociolinguistic approaches on variation (Eckert, 1989; Labov, 1963; Milroy and Milroy, 1985; Milroy and Gordon, 2003; Tagliamonte, 2006; Trudgill, 1974; Weinreich et al., 1968) work with surveys and relatively small but detailed manually collected data. Variation is analyzed considering single as well as several social variables at a time, but the small sample size affects generalization of the findings.

Increasing data availability of naturally occurring text has lead to analyze sociolinguistic variation also in corpus- and computational linguistics, especially within the social media domain (see

[1] We use stylistic variation in the sense of the workshop, i.e. variation of linguistic levels based on extra-linguistic variables (here: social variables and time).

[2] see also https://www.oldbaileyonline.org/static/Gender.jsp

e.g. Eisenstein (2015); Eisenstein et al. (2011); Nguyen et al. (2015); Danescu-Niculescu-Mizil et al. (2013); Jurafsky et al. (2009)). Recently, also the possible interplay between social variables is considered (Prabhakaran and Rambow, 2017), but is mostly confined to age and gender (see e.g. Ardehaly and Culotta (2015); Argamon et al. (2007); Barbieri (2008); Burger et al. (2011); Eckert and McConnell-Ginet (2013); Holmes and Meyerhoff (2003); Hovy and Søgaard (2015); Nguyen et al. (2014); Peersman et al. (2011); Schwartz et al. (2013); Wagner (2012)) as other social variables – such as social class – are not easily available (cf. Sloan et al. (2015)).

In fact, the gap in coverage of other social variables has recently lead to a full strand of research focusing on determining and analyzing income through Twitter content using a wide range of features. Preotiuc-Pietro et al. (2015a) use word clusters and embeddings to predict occupational class of Twitter users. Preoiuc-Pietro et al. (2015) apply non-linear methods for regression using besides shallow textual features (e.g. average no. of tweets) also user profile and psycho-demographic features (e.g. no. of followers, gender, age) as well as emotion features (e.g. positive/negative sentiments). Hasanuzzaman et al. (2017) are the first to use user cognitive structure in terms of the user's overall temporal orientation to predict income uncovering a correlation between future temporal orientation and income.

While the above mentioned literature is devoted to social media giving valuable insights into sociolinguistic, behavioral and social science research of the *present*, in this paper we study *diachronic* change in language use of social groups in approx. 200 years of court proceedings.

Considering the linguistic levels at which variation according to social variables is analyzed, in sociolinguistic approaches the phonological level prevails, while in computational approaches the lexical level is often reported to be best in prediction tasks. Other linguistic levels were mostly neglected often due to low performance of NLP tools especially for social media (e.g. sentence parsing). Recent advances in this direction have been made, for example, by Flekova et al. (2016) using besides surface features (e.g. length of tweets), readability features (such as the Automatic Readability Index or Gunning-Fog Index) as well as several style features (such as explicitness, no. of hedges) also syntax features by means of parts of speech.

In our study we are dealing with transcribed spoken utterances from the court, i.e. spoken English in a relatively formal context, thus we can consider besides lexical features also grammatical features approximated by part-of-speech trigrams. Our lexical features include content as well as function words, thus lexical as well as grammatical features will both reflect stylistic variation.

Relative entropy as a measure of divergence between corpora has been already applied successfully for the analysis of written scientific English from the 17th to the present (Degaetano-Ortlieb et al., to appear; Degaetano-Ortlieb and Teich, 2016; Degaetano-Ortlieb and Stroetgen, 2018) and for intra-textual variation, more precisely variation within sections of research articles (Degaetano-Ortlieb and Teich, 2017).

3 Data and Methods

3.1 Old Bailey Corpus

The court proceedings of the Old Bailey Court in London contain transcribed utterances of the court's trials spanning from 1674 to 1913. According to Emsley et al. (2018) the City of London "required that the publisher should provide a "true, fair and perfect narrative" of the trials" and "witness testimony is the most fully reported element of the trials". Thus, the utterances in the proceedings are arguably a relatively precise account of spoken English of that period.

The Old Bailey Corpus (OBC; Huber et al. (2016)) is built from a digitized version of the proceedings and spans from 1720 to 1913. It represents a balanced subset of the proceedings with semi-automatically identified utterances. Each utterance was semi-automatically annotated with sociolinguistic information based on sociobiographical speaker data found in the context of the trials. For this, an annotation tool was developed that first automatically detected speakers based on a list of 7,500 male and female first names (approx. 95% coverage) and in a second step allowed to scroll through the data to annotate sociobiographical information. Witnesses, for example, had to begin their statement by mentioning their profession (cf. Huber et al. (2016)). The OBC amounts at approx. 14 million spoken words (around 750,000 words per decade). It is part-of-speech tagged with CLAWS 7 (with reported accuracy of 95-98%) and sociolinguistically annotated for speaker information (gender, age, occupation according to the HISCO standard), social class (HISCLASS standard), speaker role (defendant, interpreter, judge, lawyer, victim, and witness), and textual information (scribe, printer, publisher). In addition, the corpus is divided up into years, decades and periods of fifty-years. The corpus is encoded in the Corpus Query Processor (Evert, 2005) and available for download[3] or on the CQPweb platform[4].

For the analyses, we consider the sociolinguistic annotations of gender (female, male) and social class (higher, lower) as well as the fifty-years time periods[5]. To control for speaker role, as there are no female judges or lawyers, we confine our data set to the roles of victim and witness. Table 1 gives an overview on the token size of each subcorpus.

period	FH	FL	MH	ML
1700	49,142	47,497	286,322	185,862
1750	121,942	170,090	1,084,068	855,178
1800	135,887	217,224	2,499,314	1,422,027
1850	168,246	217,830	4,069,475	1,317,113
1900	61,518	63,494	1,158,354	294,608

Table 1: Subcorpus sizes of the OBC confined to speaker role witness and victim

3.2 Detection of stylistic variation across social variables

For detecting stylistic variation, we use the method described in Fankhauser et al. (2014) based on relative entropy, precisely Kullback-Leibler Divergence (Kullback and Leibler, 1951). This approach allows us to compare probability distributions by measuring the number of additional bits needed to encode a (sub)corpus A with an optimal code for a (sub)corpus B.

$$D(A||B) = \sum_i p(unit_i|A) log_2 \frac{p(unit_i|A)}{p(unit_i|B)} \qquad (1)$$

To control for differences in vocabulary size, the corpora are represented by means of unigram language models which are smoothed with Jelinek-Mercer smoothing and lambda 0.05 (cf. Fankhauser et al. (2014) and Zhai and Lafferty (2004)).

Here, we use relative entropy to measure the difference between language use of female and male of higher and lower class over time in bits. Thus, we compare four groups (female higher class (FH), female lower class (FL), male higher class (MH) and male lower class (ML)) over five time periods (1700, 1750, 1800, 1850, 1900). For each comparison (i.e. comparison between two groups, e.g. FH vs. FL 1700, FH vs. FL 1750, etc.) a relative entropy (language) model is built. We then compare the relative entropy values obtained for each comparison to determine differences in language use across social variables and time. The higher the relative entropy value of a comparison, the more apart the two groups are and vice versa.

Note also that Kullback-Leibler Divergence is an asymmetric measure, i.e. a comparison of FH vs. FL 1700 does not necessarily result in the same relative entropy value as a comparison of FL vs. FH 1700. For comparison of variation in language use, the asymmetry is useful as it allows us to account for the directionality of the comparison.

For each comparison, we also obtain the individual unit's weight, i.e. how much a unit contributes to the difference. For example, comparing FH vs. FL 1700, we obtain the additional bits needed for a unit in FH based on the unit's probability in FL:

$$D_{unit}(FH||FL)_{1700} = p(unit|FH) log_2 \frac{p(unit|FH)}{p(unit|FL)} \qquad (2)$$

The higher the relative entropy value of a unit, the greater the unit's contribution to the difference, i.e. the more distinctive the unit is for a given group in a time period. In addition, for each comparison we test for significance of the relative frequency of a unit in the two groups by an unpaired Welch's t-test (threshold of a p-value<0.05):

$$t = \frac{mean_{FH} - mean_{FL}}{\sqrt{(\frac{var_{FH}}{n_{FH}} + \frac{var_{FL}}{n_{FL}})}} \qquad (3)$$

with var denoting the variance and n the number of documents in a group (cf. Fankhauser et al. (2014)).

To consider differences at the lexical level, the units for the relative entropy models are words. To approximate the grammatical level, we use part-of-speech (POS) trigrams as units.

In comparison to other corpus-linguistic approaches, such as classification (e.g. Teich et al.

[3] http://www1.uni-giessen.de/oldbaileycorpus/

[4] https://corpora.clarin-d.uni-saarland.de/cqpweb/usr/index.php?thisQ=accessDenied&corpusDenied=obc2&uT=y

[5] 1700: 1700-1749; 1750: 1750-1799; 1800: 1800-1849; 1850: 1850-1899; 1900: 1900-1920.

(2016)) or correspondence analysis (e.g. Glynn (2014)) just to mention a few, relative entropy directly measures the divergence between two groups in bits of information. The contribution of each unit to the divergence provides valuable insights into which units are distinctive for each group.

4 Stylistic variation across gender and social class

We investigate stylistic variation considering the interaction between gender and social class at two linguistic levels (lexis and grammar). Our focus is on change in language use of female higher class. Women's social position was increasingly questioned in the mid-nineteenth century. We hypothesize that this movement might be reflected in a change in language use of female higher class when compared to female lower class as well as male higher and lower class. As appropriate, we will also compare diachronic tendencies of the other groups.

Our concrete research questions are the following: (i) Is there a difference in language use between female higher class compared to female lower class and male higher and lower class, (ii) if so, which lexical and grammatical units contribute to these differences, (iii) do these differences change over time?

4.1 Lexical level

At the lexical level, we compare each group by relative entropy using words. From Figure 1, we can see that from 1700 to 1800 relative entropy between female higher class (FH) vs. female and male lower class (FL and ML) is lower (below 0.2 bits) than vs. male higher class (MH) (above 0.2 bits with a slight increases to 0.3 towards the period of 1800). Thus, for female higher class around 0.8 to 0.15 additional bits are needed in comparison to male higher class than from the lower class. After 1800 this changes, based on words FH becomes less distinct to MH (towards 0.2 bits), while it becomes more distinct from the lower class (vs. ML 0.2825 bits, i.e. 0.065 more bits than FHvsMH, and vs. FL 0.38 bits, i.e. 0.17 more bits than FHvsMH, in the period of 1900).

Let us compare this to male higher class (MH) vs. the other groups. From Figure 2, we see how relative entropy of MH vs. ML is relatively low (around 0.1 bits). Compared to female (FH and FL) relative entropy is higher, especially in 1700 to 1750 (i.e. around 0.2-0.25 more bits). Towards 1800 relative entropy of MH vs. FH and FL decreases. After 1800, relative entropy remains stable for MH vs. FH, while MH vs. FL increases.

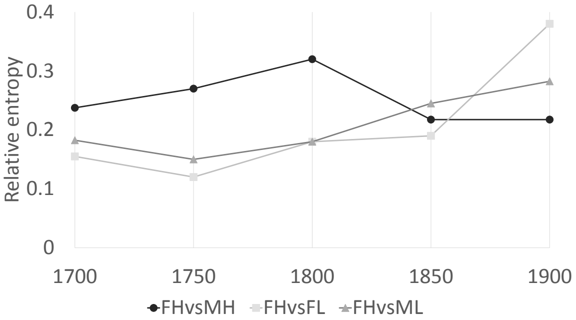

Figure 1: Relative entropy across fifty-years time periods in the OBC for female higher class (FH) vs. the other groups

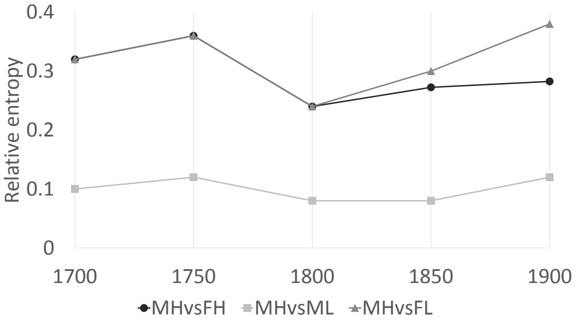

Figure 2: Relative entropy across fifty-years time periods in the OBC for male higher class (MH) vs. the other groups

Comparing Figure 1 and 2, we can see how relative entropy reflects quite well the difference related to the communicative experience of language users. Compare, for example, FH vs. MH (Figure 1) and MH vs. FH (Figure 2) in 1900 (0.2825 bits for MH vs. FH and 0.2175 bits for FH vs. MH). Here relative entropy differs due to the asymmetry of Kullback-Leibler Divergence, which allows us to model differences depending on the directionality of the comparison. Thus, if a language model of male higher class is used to predict language use of female higher class, we obtain a lower relative entropy value than vice versa. Intuitively this means that male of higher class can better understand female of higher class (here: based on words), while female of higher class need more effort (more bits) to understand male of higher class.

Let us now consider which units (here: words) contribute most to the attested differences. Consider the comparison between female higher class (FH) vs. female lower class (FL). How is the increasing difference as depicted by relative entropy (see again Figure 1) reflected in the use of words? For this, we inspect the contribution of each word to the difference (as described in Section 3.2) and visualize this in a word cloud (using the visualization approach by Fankhauser et al. (2014)). The size of the word denotes its contribution by relative entropy (in bits), the color denotes relative frequency of each word in a time period (from red for high relative frequency to blue for low relative frequency). From these clouds, we can detect variation in terms of words indicating lexical as well as stylistic differences.

As for lexical differences, FL speak distinctively about authorities (*sir, master, mistress, mr, mrs*) and objects (*door, kitchen, bedside*) related to the household; FH use distinctively business oriented vocabulary (*counter, penny, profit, purchase, business*) and words for persons related to either marriage (*husband, wife*) or crime (*officer, prisoner*).

Considering stylistic differences, while FL distinctively use personal pronouns (e.g. *his, I, me, he, him*) and verbs (e.g. *carry, become, wash, work, coming, went, going*), FH in comparison to FL over time develop a pronounced nominal style with distinctive use of nouns, definite determiners (*a, an*), and prepositions (*of, in*). Thus, female lower class use increasingly an involved verbal style over time, while female higher class make use of a nominal more informational style when compared to one another (cf. Conrad and Biber (2001, 28) for involved vs. informational production).

4.2 Grammatical level

While stylistic differences can already be seen when considering the lexical level, we consider grammatical structures approximating them by part-of-speech (POS) trigrams to detect more fine-grained tendencies. Here, we again focus on the differences between female of higher class compared to the other groups. Relative entropy models are calculated on POS trigrams as described in Section 3.2.

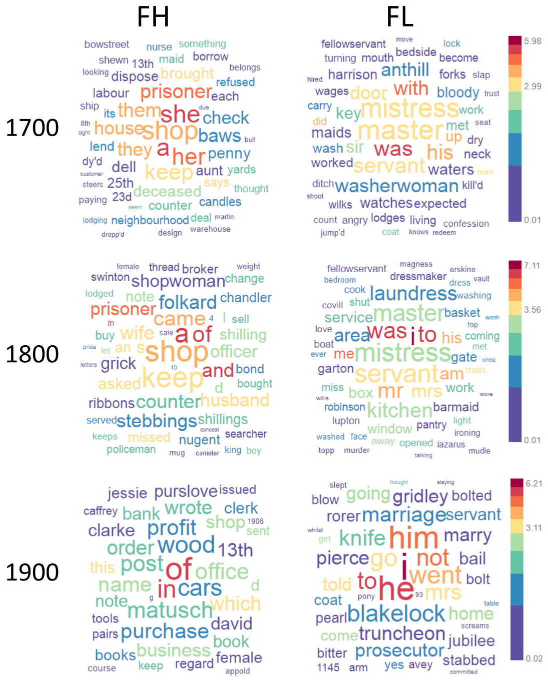

Figure 3: Words contributing to differences between female of higher (FH) vs. lower class (FL) over time (color denotes relative frequency, size relative entropy, both relative to a time period)

The greatest difference in the use of POS trigrams lies between female of higher vs. lower class with an increasing tendency over time (see FHvsFL in Figure 4), while relative entropy is lower for FH vs. male production (MH and ML). This indicates that the distribution of POS trigrams of female higher class is more similar to both male of higher and lower class than female of lower class.

type	example	bits
Female lower class		
VP (interact.)	*I keep a (House)*	0.0039
VP (interact.)	*(I) keep the Hamtshire*	0.0030
CC	*but at last*	0.0027
CC	*and found all*	0.0025
VP (interact.)	*I would have*	0.0024
Female higher class		
VP (interact.; relat.)	*(I) am Nurse at*	0.0049
NP (gen.)	*his Wife's (Clothes)*	0.0035
VP (interact.; relat.)	*I am Wife (of)*	0.0027
VP (interact.; relat.)	*(I) am the Wife (of)*	0.0025
NP+	*(to my) House from Mr.*	0.0024

Table 2: Top 5 phrase/clause types for 1700

Inspecting which POS trigrams contribute to the difference between female of higher vs. lower

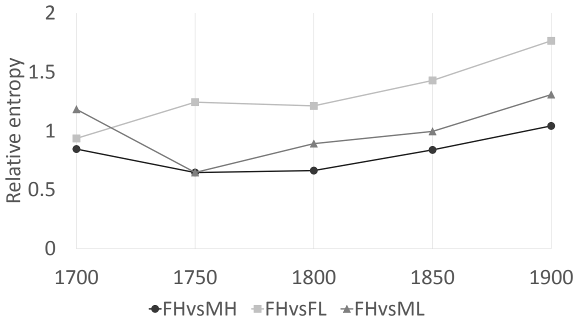

Figure 4: Relative entropy across fifty-years time periods in the OBC for females of higher class (FH) vs. male higher class (MH), female lower class (FL), and male lower class (ML)

type	example	bits
Female lower class		
VP (interact.)	*I had been*	0.0196
VP (interact.)	*asked me for*	0.0085
VP (interact.)	*said I was*	0.0080
VP (interact.)	*I could not*	0.0025
VP (interact.)	*me in the (face)*	0.0073
Female higher class		
NP+	*the intention of (committing)*	0.0304
NP	*(these are) the original invoices*	0.0208
VP (passive)	*(my attention) was directed to*	0.0165
NP+	*(contract) notes or cheques*	0.0161
VP (interact.)	*I was there (to attend)*	0.0121

Table 3: Top 5 phrase/clause types for 1900

class, we consider the contribution in bits of each POS trigram. Table 2 and 3 show the top 5 POS trigrams categorized into phrase/clause types for 1700 and 1900, respectively. In 1700, FL are distinguished from FH by a pronounced interactional style, while FH from FL by an interactional style combined with relational clauses (see example (1)). Comparing the phrase/clause types diachronically, female of higher class develop over time a nominal style (see example (2)) that distinguishes them from female of lower class, who stick to an involved verbal style (VP interact.; see examples (3) and (4)). While this is in line with the observations made at the lexical level (cf. Section 4.1), we can see which phrase/clause types are used distinctively.

(1) *I am Nurse at the Hospital; Mr. Fern examin'd the Child; she has a soul Glect, and is ulcerated in the privy Parts.* (Female higher class 1733; HISCLASS 4; HISCO label: Professional Nurse, General)

(2) *I was taken to Bow Street, where a number of people were put up for the purposes of identification. [...] In fact, I picked somebody else, a man whom I afterwards discovered to be called George Dacey.* (Female higher class 1907; HISCLASS 4; HISCO label: Mail Distribution Clerk, General)

(3) *I keep a House in Bell-Yard in King's-street, Westminster, where I sell Greens and Fruit.* (Female lower class 1734; HISCLASS 11[6]; HISCO label: Other Street Vendors, Canvassers and News Vendors))

(4) *I am a barmaid at a public-house in Tottenham on April 10th I had been out, and as I was returning home I met the prosecutor he and I and another man walked along the road together [...]* (Female lower class 1902; HISCLASS 9; HISCO label: Bartender)

POS trigrams distinctive for female higher class against all other groups are shown in Table 4 for 1700 and Table 5 for 1900. In 1700 (Table 4), interactional style is a pronounced marker of distinction, with relational clauses when compared to FL (as shown in Table 2), and with adverbial phrases and possessive phrases when compared to male production (MH and ML). Also, compared to either MH or ML, four out of five POS trigrams are identical (marked in bold). Thus, female higher class differ almost in the same way from male higher and lower class.

In 1900 (see Table 5), interactional style for FH is less distinctive (1 POS trigram compared to FL; 2 compared to MH; 1 compared to ML). Compared to FL, nominal style and passive voice are highest ranking. In comparison to both male productions (MH and ML), an adverbial/prepositional phrase and an interactional verb phrase are distinctive (marked in bold). In addition, compared to MH, a genitive noun phrase is highest ranking as well as a further adverbial phrase (AdvP). Comparison to both lower class groups (FL and ML) shows nominal style to be most distinctive: a noun phrase followed by a preposition (pointing to complex nominal phrases) is highest ranking (marked in bold).

To observe more general diachronic tendencies, we consider all top 30 POS trigrams of each comparison (i.e. for FHvsFL, FHvsMH and FHvsML). Based on the number of POS trigrams related to a

[6] 1-5 stand for higher, 6-13 for lower class.

comp.	type	POS trigram	example	bits	p-value
FHvsFL	VP (interact.) (relat.)	VB.NN1.IN	*(I) am Nurse at (the Hospital)*	0.00490	0.000107
	NP (gen.)	PP.NN1.GE	*his Wife's (Clothes)*	0.00345	0.033813
	VP (interact.) (relat.)	PPint.VB.NN1	*I am Wife (of Joseph Read)*	0.00270	0.008079
	VP (interact.) (relat.)	VB.DT.NN1	*(I) am the Wife (of Abraham Lacy)*	0.00247	0.002620
	NP+	NN1.IN.NN	*(to come to my) House from Mr. (Tull)*	0.00244	4.43E-05
FHvsMH	AdvP	**IN.PP.NN1**	*(I hid her) behind my Bed*	0.00435	0.000146
	VP (interact.) (adv.)	**PPint.VVD.RAloc**	*I came home (that Night about)*	0.00356	0.004452
	VP (interact.)	**PPint.VV0.DT**	*I keep a (Chandler's Shop)*	0.00246	0.006278
	VP (interact.)	CC.RR.PPint	*and so I (led her up stairs)*	0.00210	0.019882
	VP (interact.)	**VV0.DT.NP**	*(I) keep the Swan*	0.00196	0.020611
FHvsML	AdvP	**IN.PP.NN1**	*(I hid her) behind my Bed*	0.00731	2.60E-06
	VP (interact.)	**PPint.VV0.DT**	*I keep a (Chandler's Shop)*	0.00526	8.03E-06
	VP (interact.)	**VV0.DT.NP**	*(I) keep the Swan*	0.00349	0.002928
	VP (interact.) (poss.)	VVD.IN.PP	*(the Prisoner) came to my (House)*	0.00257	0.001892
	VP (interact.) (adv.)	**PPint.VVD.RAloc**	*I came home (that Night about)*	0.00185	0.042788

Table 4: Top 5 phrase/clause types for 1700 (overlapping POS trigrams across comparisons shown in bold)

comp.	type	POS trigram	example	bits	p-value
FHvsFL	NP+	**DT.NN1.INof**	*the intention of (committing suicide)*	0.03036	0.002835
	NP	DT.JJ.NN2	*(these are) the original invoices*	0.02075	0.028249
	VP (passive)	VBD.VVN.IN	*(my attention) was directed to (an advertisement)*	0.01649	0.035326
	NP+	NN2.CC.NN2	*(contract) notes or cheques*	0.01609	0.048210
	VP (interact.)	PPint.VBD.RAloc	*I was there (to attend)*	0.01211	0.046676
FHvsMH	AdvP/PrepP	**IN.PP.NN1**	*(this bill endorsed) by my husband*	0.01390	0.004174
	NP (gen.)	PP.NN1.GE	*my father's (banking account)*	0.00988	0.047651
	VP (interact.)	**PPint.VVD.PP**	*I saw him (sign a few letters)*	0.00887	0.020997
	AdvP	IN.DT.NPtemp	*(doing business) on a Sunday*	0.00614	0.046193
	VP (interact.)	CC.PPint.VVD	*and I made (no profit)*	0.00536	0.008897
FHvsML	NP+	**DT.NN1.INof**	*the intention of (committing suicide)*	0.01936	0.001253
	VP (interact.)	**PPint.VVD.PP**	*I saw him (sign a few letters)*	0.01069	0.002745
	NP+	NN1.INof.NN1	*(The) consignment of paper (came during)*	0.00830	0.020229
	AdvP/PrepP	**IN.PP.NN1**	*(this bill endorsed) by my husband*	0.00666	0.025071
	VP+	VVD.PP.IN	*(she) sent it to (me from Ostend)*	0.00365	0.015710

Table 5: Top 5 phrase/clause types for 1900 (overlapping POS trigrams across comparisons shown in bold)

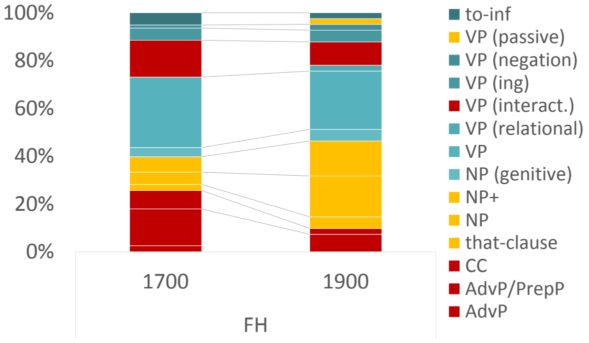

Figure 5: Percentage of top 30 POS trigrams by phrase/clause types distinctive for female of higher class (FH)

phrase type, we calculate the percentage of each phrase type distinctive of female higher class for both time periods (see Figure 5[7]). Red denotes phrase types which become less distinctive over time, yellow denotes phrase types more distinctive over time. While interactant verb phrases (VP interact.) become less distinctive for female higher class, nominal phrase types (NP genitive, NP+, NP) are considerably more distinctive over time. Phrases with conjunctions (CC; e.g. *he got up, and came to me*) as well as adverbial and prepositional phrases are less distinctive over time. The percentage of nominal style distinctive for female higher class increases over time (from 15% to 37%), while a distinctive verbal style decreases (from 56% to 49%), especially an interactant verbal style (from 15% to 9.8%) .

[7] From top to bottom: to-infinitives (to-inf), passive voice (VP passive), negation (VP negation), -ing form (VP ing) interactant verb phrases (VP interact.), simple verb phrase (VP), genitives (NP genitive), complex noun phrases (NP+, i.e. with prepositions or coordinative conjunctions), and simple noun phrases (NP), conjunctions (CC), adverbial and prepositional phrases (AdvP/PrepP), adverbial phrases of degree, location, comparison etc. based on the CLAWS7 tag set.

5 Conclusion

We have presented an approach to investigate stylistic variation across social variables and time at two linguistic levels: lexis and grammar. Our focus was on language use of female of higher class in court proceedings over the time span of approx. 200 years. We asked whether the uprising feminist movement from the mid-nineteenth century, questioning the women's inferior social position, is reflected in a change in language use of female higher class.

In terms of methods, we have used relative entropy according to Fankhauser et al. (2014), which allows us to measure the difference between probability distributions of linguistic units (here: words and POS trigrams). At the lexical level, lexical as well as stylistic differences have been identified. At the grammatical level, more fine-grained stylistic differences have been detected: female of higher class developed over time a nominal more informational style that increasingly differs from female of lower class.

While we have focused on female of higher class, in our ongoing work, we are analyzing the development of each group. Moreover, we will also consider the other roles in the trials, which will give more detailed insights into the development of language use in court trials. Also, while we use social class distinction based on higher and lower class, a more fine-grained distinction could be used as the OBC is annotated on a scale from 1-13. Instead of considering fifty-years time periods for comparison, in future work we aim to detect in which time span a particular change takes place.

In terms of contributions, by using an approach based on information theory (i.e. relative entropy), we are able to model language use and directly compare different groups of language users with one another, also obtaining linguistic units distinctively used across groups. The models are based on whole linguistic levels rather than on predefined features. This allows for a systematic account of language/stylistic variation. Also, we have shown that stylistic variation of groups may well change over time.

Acknowledgments

This research is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SFB1102: Information Density and Linguistic Encoding (www.sfb1102.uni-saarland.de) and the start-up grant for research projects from Saarland University. I would also like to thank the anonymous reviewers for their constructive and valuable comments, Peter Fankhauser (IDS Mannheim) for support in questions of data analysis, Stefan Fisher (Saarland University) for corpus processing, Magnus Huber (Giessen University) for questions on the OBC, and Elke Teich (Saarland University) for valuable comments on the first draft of this paper.

References

Ehsan Mohammady Ardehaly and Aron Culotta. 2015. Inferring Latent Attributes of Twitter Users with Label Regularization. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 185–195, Denver, Colorado.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. Mining the Blogosphere: Age, Gender and the Varieties of Self-expression. *First Monday*, 12(9).

Federica Barbieri. 2008. Patterns of Age-based Linguistic Variation in American English. *Journal of Sociolinguistics*, 12(1):58–88.

Basil Bernstein. 1971. *Class, Code and Control: Volume 1 Theoretical Studies towards a Sociology of Language*. Routledge Taylor & Francis Group.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh, Scotland, UK.

Susan Conrad and Douglas Biber. 2001. *Variation in English: Multi-Dimensional Studies*. Longman, London.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 250–259, Sofia, Bulgaria. ALC.

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, and Elke Teich. to appear. *From Data to Evidence in English Language Research*, Language and Computers, chapter An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English. Brill.

Stefania Degaetano-Ortlieb and Jannik Stroetgen. 2018. Diachronic Variation of Temporal Expressions in Scientific Writing Through the Lens of Relative Entropy. In *Language Technologies for the*

Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings, volume 10713 of *Lecture Notes in Computer Science*, pages 259–275. Springer International Publishing.

Stefania Degaetano-Ortlieb and Elke Teich. 2016. Information-based Modeling of Diachronic Linguistic Change: From Typicality to Productivity. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Berlin, Germany. ACL.

Stefania Degaetano-Ortlieb and Elke Teich. 2017. Modeling Intra-textual Variation with Entropy and Surprisal: Topical vs. Stylistic Patterns. In *Proceedings of the Joint LaTeCH and CLfL Workshop at ACL*, pages 68–77, Vancouver, Canada. ACL.

Penelope Eckert. 1989. *Social Categories and Identity in the High School*. Teachers College Press.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and Gender*. Cambridge University Press, Cambridge.

Jacob Eisenstein. 2015. Written Dialect Variation in Online Social Media. In Charles Boberg, John Nerbonne, and Dom Watt, editors, *Handbook of Dialectology*. Wiley.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011. Discovering Sociolinguistic Associations with Structured Sparsity. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1365–1374, Portland, OR.

Clive Emsley, Tim Hitchcock, and Robert Shoemaker. 2018. The Proceedings - The Value Of the Proceedings as a Historical Source. Old Bailey Proceedings Online, version 7.0.

Stefan Evert. 2005. *The CQP Query Language Tutorial*. IMS Stuttgart. CWB version 2.2.b90.

Peter Fankhauser, Jörg Knappen, and Elke Teich. 2014. Exploring and Visualizing Variation in Language Resources. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

Lucie Flekova, Daniel Preotiuc-Pietro, and Lyle H. Ungar. 2016. Exploring Stylistic Variation with Age and Income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 313–319, Berlin, Germany. ACL.

Dylan Glynn. 2014. Correspondence Analysis - Exploring Data and Identifying Patterns. In Dylan Glynn and Justyna A. Robinson, editors, *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, volume 43 of *Human Cognitive Processing*, pages 443–485. John Benjamins, Amsterdam.

Mohammed Hasanuzzaman, Sabyasachi Kamila, Mandeep Kaur, Sriparna Saha, and Asif Ekbal. 2017. Temporal Orientation of Tweets for Predicting Income of Users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 659–665. Association for Computational Linguistics.

Janet Holmes and Miriam Meyerhoff. 2003. *The Handbook of Language and Gender*. Wiley-Blackwell.

Dirk Hovy and Anders Søgaard. 2015. Tagging Performance Correlates with Author Age. In *Proceedings of the 53rd AnnualMeeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 2, pages 483–488, Beijing, China. ACL.

Magnus Huber, Magnus Nissel, and Karin Puga. 2016. *Old Bailey Corpus 2.0*. Hdl:11858/00-246C-0000-0023-8CFB-2.

Dan Jurafsky, Rajesh Ranganath, and Dan McFarland. 2009. Extracting Social Meaning: Identifying Interactional Style in Spoken Conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 638–646, Boulder, Colorado.

Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

William Labov. 1963. The Social Motivation of a Sound Change. *Word*, 19(3):273–309.

James Milroy and Lesley Milroy. 1985. Linguistic Change, Social Network and Speaker Innovation. *Journal of Linguistics*, 21(2):339–384.

Lesley Milroy and Matthew Gordon. 2003. *Sociolinguistics: Method and Interpretation*. Wiley-Blackwell.

Dong Nguyen, A. Seza Dogruöz, Carolyn Penstein Rosé, and Franciska de Jong. 2016. Computational Sociolinguistics: A Survey. *CoRR*, abs/1508.07544.

Dong Nguyen, Dolf Trieschnigg, and Leonie Cornips. 2015. Audience and the Use of Minority Languages on Twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 666–669, Oxford, UK.

Dong Nguyen, Dolf Trieschnigg, A. Seza Dogruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. 2014. Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, Dublin, Ireland.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting Age and Gender in Online Social Networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, pages 37–44, New York, NY, USA. ACM.

Vinodkumar Prabhakaran and Owen Rambow. 2017. Dialog Structure Through the Lens of Gender, Gender Environment, and Power. *Dialogue & Discourse*, 8(2):21–55.

Daniel Preoiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying User Income through Language, Behaviour and Affect in Social Media. *PLOS ONE*, 10(9):1–17.

Daniel Preotiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015a. An Analysis of the User Occupational Class through Twitter Content. In *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of Natural Language Processing*, pages 1754–1764, Beijing, China. ACL.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE*, 8(9):1–16.

Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. 2015. Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. *PLOS ONE*, 10(3):1–20.

Sali A. Tagliamonte. 2006. *Analysing Sociolinguistic Variation*. Cambridge University Press, Cambridge.

Elke Teich, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes, and Ekaterina Lapshinova-Koltunski. 2016. The Linguistic Construal of Disciplinarity: A Data-mining Approach Using Register Features. *Journal of the Association for Information Science and Technology (JAIST)*, 67(7):1668–1678.

Peter Trudgill. 1974. *The Social Differentiation of English in Norwich*. Cambridge University Press, Cambridge.

Suzanne E. Wagner. 2012. Age Grading in Sociolinguistic Theory. *Language and Linguistics Compass*, 6(6):371–382.

Uriel Weinreich, William Labov, and Marvin I. Herzog. 1968. Empirical Foundations for a Theory of Language Change. In Winfred P. Lehmann and Yakov Malkiel, editors, *Directions for Historical Linguistics: A Symposium*, pages 95–188. University of Texas Press, Austin.

Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, 22(2):179–214.