



RESEARCH ARTICLE

Exploring the language of Swedish social media: A contrastive corpus analysis

Evie Coussé¹  and Yvonne Adesam² 

¹Department of Languages and Literatures, University of Gothenburg, Box 200, 405 30 Göteborg, Sweden and ²Department of Swedish, Multilingualism, Language Technology, Språkbanken Text, University of Gothenburg, Box 200, 405 30 Göteborg, Sweden

Corresponding author: Evie Coussé; Email: evie.cousse@gu.se

(Received 3 May 2024; revised 18 December 2024; accepted 19 December 2024)

Abstract

This article explores the language of social media by analyzing a selection of linguistic features in four corpora of Swedish social media available at Språkbanken Text: Blog mix, Familjeliv, Flashback, and Twitter. Previous research describes the language of these corpora as informal, spoken-like, unedited, non-standard, and innovative. Our corpus analysis confirms the informal and spoken-like nature of social media, while also showing that these traits are unevenly distributed across the various social media corpora and that they are also present in other traditional written corpora, such as novels. Our findings also reveal that the social media corpora show traits of involved and interactional language.

Keywords: blog; corpus; discussion forum; informal; social media; spoken; Swedish; Twitter; written

1. Introduction

The advent of computer-mediated communication has made available for corpus linguistics an increasingly growing and diverse body of written language for investigation. In this article, we investigate social media as a source for corpus linguistic research. Social media include a wide variety of online platforms and websites ‘that have been purpose-built to facilitate social interaction and that rely on user content to fill their pages’ (Rüdiger & Dayter 2020:2).

We focus on Swedish social media corpora available from Språkbanken Text.¹ Blog mix, the oldest social media corpus at Språkbanken Text, first released in 2013, contains a collection of the most popular blogs in Sweden. Twitter is a corpus of Swedish microblogs from the platform Twitter, which was rebranded as X in 2023. Familjeliv and Flashback are corpora with posts from two Swedish discussion forums. All of these corpora have in common that they are very large (billions of words altogether) and contain data from the last twenty-five years (from the period 1998–2024).

© The Author(s), 2025. Published by Cambridge University Press on behalf of The Nordic Association of Linguists. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Språkbanken's social media corpora have been a popular source for corpus studies of contemporary Swedish (e.g. Ledin & Lyngfelt 2013, Olofsson 2014, and Bylin 2016, to name only a few early examples). These studies describe the language of social media as being informal, spoken-like, unedited, non-standard, and innovative. These properties are not systematically and empirically assessed in these studies but are rather provided as part of a rationale for choosing social media as a source for investigation, often in contrast with or in addition to more traditional language sources. This article aims to provide an empirical underpinning of these properties in social media corpora. We do this by means of a quantitative corpus study of social media corpora in contrast with more traditional corpora of written language. This comparative perspective allows us to explore the differences and similarities between the language of social media and other types of written language in Swedish.

Our study is a follow-up to Wiktorsson (2018), a short communication published in this journal, which investigates some linguistic features of the Swedish Blog mix corpus. Since Wiktorsson's article, more social media corpora have become readily available at Språkbanken Text. We are therefore in the position to explore the language of Swedish social media from a wider perspective. Our study not only covers more types of social media but also explores a wider range of properties at different linguistic levels. Wiktorsson's focus is on examining whether blogs can be considered a hybrid of spoken and written language. Our focus is not only on the spoken-like properties of social media, but also on their informal, unedited, and non-standard nature.

The structure of our article is as follows. We first give an overview of properties discussed in previous corpus studies of Swedish social media (Section 2.1), followed by a literature review of international studies looking into register variation in computer-mediated communication and social media (Section 2.2). This literature overview forms the basis for selecting linguistic features that will be used to explore the language of social media in Swedish (Section 2.3). We then describe the corpora and method used in this study in more detail (Section 3). Section 4 presents a contrastive analysis of the linguistic features in our corpora. We close with a short conclusion in Section 5.

2. Background

2.1 Linguistic corpus studies of Swedish social media

The social media corpora of Språkbanken have been used in several corpus studies of Swedish. Many of these studies motivate their use of social media corpora, as an alternative or in addition to more traditional language sources, by highlighting some of their unique properties. We were able to identify five recurring properties of social media in a sample of 54 corpus studies.² Table 1 gives an overview of these properties, along with alternative terms used in the studies to describe them, and the corpus studies that make explicit mention of them.

Table 1 shows that the most mentioned property of social media is its informal character. Brandtler (2019), for instance, argues that the language of Flashback 'tends to reflect a natural and informal style of speech' which is assumed to be

Table 1. Properties of social media in corpus studies of Swedish

Properties	Alternative descriptions in corpus studies	Corpus studies	n
Informal	Casual, colloquial, everyday, intimate, natural	Blensenius 2013, Julien & Lødrup 2013, Ledin & Lyngfelt 2013, Ahlberg et al. 2015, Engdahl & Laanemets 2015, Hillbom 2015, Bylin 2016, Jansson 2016, Malm 2016, Åkerblom 2016, Engdahl & Coppock 2017, Sköldberg & Hannesdottir 2017, Brandtler 2019, 2020, Caplan & Djärv 2019, Olofsson & Prentice 2020, Thyberg 2020, Valdeson 2021, Berdicevskis, Adesam & Coussé 2022, Collberg & Agebjörn 2022, Katourgi 2022, Wiktorsson 2022, Coussé et al. 2023, Höder 2023, Adesam, Berdicevskis & Coussé 2024	25
Spoken-like	Close to spoken language, resembles spoken dialogue, with spoken features, a hybrid between spoken and written language	Skärllund 2014, Ahlberg et al. 2015, Hillbom 2015, Rawoens 2015, Åkerblom 2016, Caplan & Djärv 2019, Berger 2020, Blensenius & Rogström 2020, Brandtler 2020, Olofsson & Prentice 2020, Valdeson 2021, Wiktorsson 2022, Höder 2023, Adesam, Berdicevskis & Coussé 2024	14
Unedited	Spontaneous, minimum of text planning	Ahlberg et al. 2015, Malm 2016, Brandtler 2019, 2020; Blensenius & Rogström 2020, Berdicevskis, Adesam & Coussé 2022, Katourgi 2022, Coussé et al. 2023, Adesam, Berdicevskis & Coussé 2024	9
Innovative	Shows innovations earlier, innovations spread faster	Ledin & Lyngfelt 2013, Hillbom 2015, Skärllund 2016, Blensenius & Rogström 2020, Berdicevskis, Adesam & Coussé 2022, Coussé et al. 2023, Adesam, Berdicevskis & Coussé 2024	7
Non-standard	Follows norms and conventions of the written standard language to a lesser extent	Ahlberg et al. 2015, Valdeson 2017, Brandtler 2020, Berdicevskis et al. 2022, Coussé et al. 2023	5

‘representative of modern-day, informal Swedish’ (Brandtler 2019:762). Bylin (2016:81), in turn, takes Familjeliv to represent everyday private language in contrast to the public language of news texts.

Another frequently highlighted property of the language of social media is its spoken-like nature, the property in focus in Wiktorsson (2018). Blensenius & Rogström (2020), more specifically, point out that social media language often has a ‘momentary and sometimes dialogical character’ (Blensenius & Rogström 2020:84). Caplan & Djärv (2019) focus on the discussion forum Flashback, stating that it ‘relates more closely to spoken dialogue compared with some of the other written material’ (Caplan & Djärv 2019:13).

The unedited nature of social media is also often mentioned. Brandtler (2019) highlights the non-planned nature of the discussion forum Flashback, where linguistic choices are less likely to result 'from deliberate attempts by the writer to achieve literary or rhetorical effects compared to more self-conscious genres, such as fiction and blogs' (Brandtler 2019:762). Malm (2016:16), similarly, contrasts the spontaneous language of Familjeliv with that of more heavily edited literary works or administrative texts.

Social media language is also considered to be less standard than other types of written language. Brandtler (2020:66), more specifically, expects the discussion forum Familjeliv to follow language norms to a lesser degree than the language of news and novels. The non-normative characteristics of social media language are often mentioned together with the unedited characteristics. Berdicevskis, Adesam & Coussé (2022) write, for instance, that the language of Flashback 'is closer to everyday use and less edited than many other types of text, which may mean that non-normative language is more frequent or shows up earlier than in edited texts' (Berdicevskis, Adesam & Coussé 2022:6).

All of these linguistic properties are in turn connected to the innovativeness of social media. Berdicevskis, Adesam & Coussé (2022) and Adesam, Berdicevskis & Coussé (2024) state that social media incorporate innovation faster than, for instance, news texts. Other studies that focus on the quick spread of innovations in social media are Ledin & Lyngfelt (2013:148), who suggest that the trendy and everyday language of blogs allows for fast incorporation of innovations like the new Swedish gender-neutral pronoun *hen*, and Blensenius & Rogström (2020:85), who find fast changes in the grammatical structure of multi-word expressions in social media.

While the above corpus studies contribute to identifying relevant properties of social media in Swedish, they do not assess them empirically. It is our aim to substantiate the above properties by means of a quantitative corpus study of social media corpora in contrast with other more traditional linguistic corpora. This comparative perspective will enable us to uncover the differences and similarities between social media and other corpora of Swedish.

2.2 Register variation studies of computer-mediated communication

Our corpus-based contrastive study of social media draws on international research on computer-mediated communication from a text-linguistic point of view. Computer-mediated communication (CMC) comprises more types of digital texts than only social media, and is defined by Herring (1996) as 'communication that takes place between human beings via the instrumentality of computers' (Herring 1996:1).

One of our major sources of inspiration is corpus linguistic studies contrasting the language of one form of computer-mediated communication with that of other spoken and written corpora (Yates 1996, Radić-Bojanić 2006, Jensen 2007, Ling & Baron 2007, Tagliamonte & Denis 2008, Wiktorsson 2018). The classical CMC study of Yates (1996) investigates a few linguistic features in a sample of discussions taken from early computer referencing systems (predecessors of the contemporary discussion forums) in contrast to a corpus of spoken and written language. The

study is framed in the early functional grammar of Halliday (1978), differentiating between the textual, interpersonal, and ideational aspects of language use. This framework posits that language use is variable and may vary from one situation to another, thereby exhibiting register variation. Register is here defined as ‘the clustering of semantic features according to situation type’ (Halliday 1978:111). This particular theory of register, and its later developments (Halliday 1985), has played an influential role in the text-linguistic study of computer-mediated communication and also forms the overall theoretical background of our study.

Yates first explores textual metrics such as type/token ratio and lexical density. Lexical density gives an insight into the proportion of lexical versus functional words in a corpus. He finds that the lexical density of spoken language is lower than for written language, confirming previous research of Ure (1971) and Halliday (1985), and that computer referencing is closer to written than spoken language in this regard. Yates also investigates the interpersonal aspect of language by comparing the frequency distribution of personal pronouns and modals. His study of pronouns draws on the work of Chafe (1982), who relates the use of first-person reference to the speaker’s involvement with his or her audience. Yates finds that the relative frequency of first-person pronouns is higher in spoken than in written language, confirming Chafe’s findings (1982), and that computer referencing is closer to spoken language in its use of first-person pronouns. Overall, Yates’ contrastive corpus study demonstrates that only a few linguistic features allow us to differentiate computer-mediated communication from other registers. It turns out that online language is more similar to spoken language with regard to some features and more similar to written language for other features.

Wiktorsson’s (2018) study of Swedish blogs can be placed in the tradition of Yates (1996) and subsequent contrastive corpus studies. Wiktorsson compares three frequency measures extracted from the Blog mix corpus with frequency data reported by Allwood (1998) for a corpus of spoken and written Swedish. The lexical diversity of the blog data, as measured by type/token ratio, turns out to be in between the ratios reported for speech and writing by Allwood (1998). The vocabulary variance of blogs, as measured by frequency-based rank, appears to be closer to writing than to speech. An investigation of the ten most frequent words shows that blog texts ‘are more like writing from a basic structural perspective, but perhaps display certain personal involvement features normally associated with interactive speech’ (Wiktorsson 2018:375). Similarly to Yates (1996), Wiktorsson concludes that blogs take a unique position in relation to the other language varieties investigated: sometimes the blog data shares features with spoken language; at other times it is more like written language.

A more recent development in the quantitative corpus-based study of social media is the application of multidimensional analysis (MD) ‘to empirically analyze the ways in which linguistic features co-occur in texts and the ways in which registers vary with respect to those co-occurrence patterns’ (Biber 2019:49). This approach originates in the seminal work of (Biber 1988) on spoken and written registers of English and has developed into a well-established quantitative framework for the study of register variation. The focus on studying co-occurring linguistic features goes beyond the corpus studies described earlier, such as Yates (1996) and Wiktorsson (2018), which chart the distribution of linguistic

features separately. The MD method has been applied on a large scale to online texts by Biber & Egbert (2018) and to various types of social media by Berber Sardinha (2014, 2018, 2022), Friginal, Waugh & Titak (2018), Liimatta (2019, 2023), Clarke (2019, 2022), and Clarke & Grieve (2019).

The multidimensional analysis of social media is a powerful tool that allows the uncovering of situational dimensions that correspond to the major communicative functions of these texts. Berber Sardinha (2022) identifies two situational dimensions for social media in English (Facebook, Instagram, Reddit, Telegram, Twitter, and YouTube). The first dimension 'is packed with features that enable a formal type of posting that generally conveys planned, edited, highly informational and argumentative content' (Berber Sardinha 2022:663). The set of features in the second dimension 'indicate informality, speaker orientation, engagement, and interaction' (Berber Sardinha 2022:668). What is interesting is that these dimensions go beyond the classical distinction between spoken and written language that has dominated register variation studies (see Biber 2014 for an overview). Other properties are highlighted which have also been mentioned in the corpus studies of Swedish, such as the degree of formality or editing.

Multidimensional analysis has, to the best of our knowledge, not been applied to Swedish data (social media or other). Indeed, as Goulart & Wood's (2019:129) literature review indicates, the majority of MD register studies are based on English. We see a number of reasons why this might be the case. First, multidimensional analysis relies on frequency information on dozens of lexico-grammatical features that are known to be sensitive to register variation. While there is a long research tradition in English that offers such background information, this is not the case for Swedish. Second, the automatic annotation of these lexico-grammatical features requires a purpose-made tagger identifying a large number of word classes complemented with semantic information. Again, while such taggers are available for English – notably the Biber-tagger – they are not for Swedish. There are linguistic annotations available for the social media corpora in Språkbanken that could provide a good starting point for building such a tagger, but here we should keep in mind that these annotations are not as reliable for social media corpora as for corpora with standard language (see e.g. Adesam & Berdicevskis 2021). In view of these obstacles, we refrain from performing a multidimensional analysis of our social media data, but rather explore a selection of linguistic features separately, building on the tradition set in early contrastive corpus studies of computer-mediated communication.

What differentiates our research from these pioneering studies is that we move beyond comparing two or three small data samples of a few ten thousand words to comparing half a dozen corpora consisting of millions of words (see Table 2). This trend towards 'ever-growing corpora' in the study of social media 'became an extra stimulus for computational linguistics and statistical data processing' (Vandekerckhove et al. 2019:158). Some of these computational linguistic studies of social media have been a direct inspiration for our study. Hu, Talamadupula & Kambhampati (2013) situate the language of Twitter in the spectrum of other written registers of English (SMS, Chat, Email, Blog, Online Magazines, and News) by exploring a few classical lexico-grammatical features. They find that 'the language of Twitter is highly dynamic, and that depending on the measure that is used, it

Table 2. Basic statistics of corpora under investigation

Corpus	Subcorpora	#	Period	Last updated	Tokens (M)
Blog mix	Blog mix <i>year</i> ^a	20	1998–2017	2017	581
Familjeliv	Subforums	23	2003–2023	2023	4,497
Flashback	Subforums	16	2000–2023	2024	4,533
Twitter	Twitter mix	1	2006–2022	2022	2,112
	Swedish Twitter <i>year</i>	3	2015–2017	2018	
News	Göteborgs-Posten <i>year</i>	14	1994, 2001–2013	2017	795
	Göteborgs-Posten två dagar	1	n.d.	2017	
	SVT news <i>year</i>	20	2004–2023	2022–2023	
	Dagens Arena	1	2007–2023	2024	
	Web news <i>year</i>	13	2007–2013	2022–2024	
Novels	Bonniers I	1	1976–1977	2017	13
	Bonniers II	1	1980–1981	2017	
	Norstedts	1	1999	2017	
Wikipedia		1	2013–2022	2023	190

^aItalicized *year* is used as a placeholder for the years listed in the column ‘Period’. The subcorpora listed for Blog mix should thus read as ‘Blog mix 1998’, ‘Blog mix 1999’, etc. until ‘Blog mix 2017’.

shows similarities to different media’ (Hu, Talamadupula & Kambhampati 2013:244). This study demonstrates that social media can insightfully be contrasted with other types of written language, going beyond the classical dichotomy of spoken and written language. Berdicevskis (2013) compares fewer registers, notably Russian emails and chat messages, but extends the range of linguistic features from the traditional set of lexico-grammatical features to other features, such as the mean length of an utterance or the use of punctuation. These other features have the advantage that they are relatively easy to extract automatically from a large corpus, without relying on advanced annotations. Ortmann & Dipper (2019) further contribute to expanding the set of ‘linguistic features that (i) are useful predictors of the conceptual orality of a given text and (ii) can be recognized fully automatically in texts of any length’ (Ortmann & Dipper 2019:66) in their research on German registers. The research of Berdicevskis (2013) and Ortmann & Dipper (2019) has inspired us to broaden the set of linguistic features explored for Swedish. These studies also show that linguistic features proposed in the international literature can be successfully applied to languages other than English.

2.3 Selection of linguistic features

The literature review revealed that there are dozens of potential linguistic features that have not yet been explored for Swedish. We selected from the literature a handful of features guided by the following principles. First, we only selected

linguistic features that could be extracted automatically from our corpora with good confidence, that is, frequency information based on counting particular word forms or parts of speech. Second, we chose a few features that are well-established in the early literature on computer-mediated communication and complemented these with some features suggested in more recent computational linguistic research. As such, we aimed to expand the range of linguistic features explored for Swedish. Third, we selected features for which we may assume that they are sensitive to one or more of the properties highlighted in the corpus studies of Swedish social media, that is, informal, spoken-like, unedited, and non-standard. Note that the link between linguistic features and the properties to which they may point is still tentative for Swedish. As we mentioned earlier, most of the literature on register variation has been focused on English, so it remains to be seen whether the features selected from the international literature are helpful in differentiating corpora in Swedish.

2.3.1 *Personal pronouns*

One of the classical features in register variation research is personal pronouns. Numerous corpus-based register studies have shown that the frequency and distribution of first-, second-, and third-person pronouns differ in spoken and written language (see e.g. Chafe 1982, Biber 1986, 1988, Yates 1996, and Tagliamonte & Denis 2008 for English). These studies systematically reveal that first- and second-person pronouns are more frequent in spoken language, whereas third-person pronouns are predominant in written language. Chafe (1982) relates the use of first- and second-person pronouns to the speaker's involvement with his or her audience. Biber (1988:225) similarly considers first-person pronouns as markers of ego-involvement and interpersonal focus. Some corpus studies compare the use of personal pronouns in computer-mediated language with that found in spoken and written language. Yates (1996), Radić-Bojanić (2006), and Tagliamonte & Denis (2008) find that the language of computer conferencing, chat, and instant messaging has a high relative usage of first- and second-person pronouns, which makes it similar to speech. Hu, Talamadupula & Kambhampati (2013) compare several types of computer-mediated language, finding a more diverse distribution of personal pronouns: for instance, a high use of first- and second-person pronouns on Twitter and in chats, and a high usage of third-person pronouns in blogs. The research overall suggests that personal pronouns have a good potential to separate our corpora with regard to their degree of involvement versus detachment.

2.3.2 *Nominal style and lexical density*

Two other well-known features in corpus-based register studies target the distribution of word classes. The first feature involves the relative proportion of verbs and nouns. Spoken language is reported to have a more verbal style with a higher proportion of verbs, whereas written language has a more nominal style with a higher number of nouns and nominalizations (Chafe 1982, Ortmann & Dipper 2019). Chafe (1982) relates a high degree of nominalization in written language to integration, i.e. 'the packing of more information into an idea unit than the rapid

pace of spoken language would normally allow' (Chafe 1982:39). Ortmann & Dipper (2019), similarly, interpret nominalizations as a sign of high complexity. Chafe (1982) further suggests that nominalization can function not only as an integrative device but also allows a statement to be presented in a more abstract, detached way. Radić-Bojanić (2006) finds that the discourse of electronic chat rooms contains very few nominalizations, and as such resembles spoken language more than written language.

The second feature is lexical density, that is, the proportion of lexical words versus function words. Speech is shown to have a lower lexical density than written language (Ure 1971, Halliday 1985, Yates 1996). Yates (1996) finds that the lexical density of computer conferencing is closer to that of written language than of speech. Hu, Talamadupula & Kambhampati (2013) report that, among several computer-mediated communication forms, SMS and chat have the lowest lexical density, blogs and news the highest, with Twitter taking an in-between position. They suggest that the length restriction on Twitter messages influences their lexical density. Initially, Twitter had a length restriction of 140 characters per post, which was then doubled to 280 characters in November 2017. On the basis of the above research, we assume that nominal style and lexical density have the potential to differentiate between our corpora.

2.3.3 Interjections

A feature that is less explored in corpus-based register studies is the use of interjections. A few studies have shown that they are more frequent in spoken than in written language (Allwood 1998 for Swedish, Ortmann & Dipper 2019 for German), which signals that they are sensitive to register variation. Interjections are of interest to us as they are reported to be associated with varying stylistic values in Swedish (Teleman et al. 1999). They thus have the potential to offer insights into the degree of formality of our corpora.

We select a few frequent interjections specialized in answering with different stylistic values; more specifically, the answer interjections *ja* 'yes', *jo* 'yes' (to disagree with a negated statement), *nej* 'no', and *nä* 'no'. The particles *ja*, *jo*, and *nej* are relatively neutral in style. The form *nä*, a frequent spelling variant of *nej*, has a casual, informal stylistic value, as Teleman et al. (1999:754) point out. Answer interjections are often used to react to the preceding utterance of the interlocutor (Teleman et al. 1999:756–758). As such, they may be expected to be indicative of interactional language.

We also look into some emotional interjections, which express an emotional reaction on the part of the speaker, often to what has been said before. Teleman et al. (1999:748–750) suggest that they have an informal stylistic value and are rarely used in fact-based written language. Additionally, these emotional interjections may be expected to be indicative of an involved style, as they express the personal emotions of the speaker. We selected both the stylistically neutral interjections *oj* 'oh, oops' and *usch* 'ugh, ew' and the profanities *helvete* 'damn' (lit. 'hell'), *jävlar* 'damn' (lit. 'devils'), and *herregud* 'oh my god' (lit. 'lord god') with a lower stylistic value. Note that the profanities are so-called secondary interjections (Ameka 1992), which may also function as regular nouns.

2.3.4 *Sentence punctuation*

A feature that goes beyond the classical lexico-grammatical features is the use of sentence punctuation. Ortman & Dipper (2019:69) point out that there is more variation in sentence types in spoken than written language; more specifically, ‘questions and exclamations are more frequent in spoken than in written language’ (Ortman & Dipper 2019:69). This finding suggests that sentence punctuation can contribute to differentiating our corpora. We particularly look into the use of the period, question mark, and exclamation mark to mark the end of sentences. We also report on the repeated use of punctuation marks in these contexts, which is considered a clear indicator of conceptual orality by Ortman & Dipper (2019). We hypothesize that it might also point to non-standard language use.

2.3.5 *Sentence length*

Our final feature targets the length of sentences. Ortman & Dipper (2019) find that spontaneous spoken communication and online chat, which they consider oral-oriented registers, have a systematically lower average sentence length than rehearsed TED talks, recited speeches, and newspaper articles, which are literate-oriented registers in their view. We will explore whether sentence length can help us differentiate between our corpora and uncover differences in their degree of orality.

3. Corpora and method

We now present the corpora under investigation in more detail. We use the four large social media corpora available at Språkbanken Text: Blog mix, Familjeliv, Flashback, and Twitter. We complement these corpora with three other corpora of written texts from the same corpus infrastructure: Novels, News, and Wikipedia.³ Novels and news are well-established data sources in Swedish linguistics and are often taken to be representative of contemporary standard language in its written form (Allén 1970, Ledin & Lyngfelt 2013). We also include Wikipedia, which is considered a type of social media by some (Lomborg 2011, Deumert 2016, Aichner et al. 2021), but is not included in the list of social media corpora of Språkbanken. This corpus is generally overlooked in corpus studies of Swedish. The basic statistics of the seven corpora are presented in Table 2.

We have extracted frequency information on the linguistic features to be analyzed in Section 4 via the corpus search platform Korp (Borin, Forsberg & Roxendal 2012), either through the API⁴ or through the word statistics files⁵ available for each corpus. Our extraction of the selected linguistic features relies as little as possible on the automatic annotations provided in Korp, which may be less reliable for our social media corpora, as noted earlier. Personal pronouns and interjections are therefore searched for on the basis of their word forms alone. Lexical density makes use of part-of-speech annotations, which may contain errors for individual words but are relatively reliable on the aggregate level. Sentence punctuation and sentence length rely most heavily on automatic annotations, more specifically, tokenization and sentence segmentation. We want to highlight that there may be segmentation errors, in particular in the social media corpora, as Korp’s annotation tools have difficulties

handling the various smileys and other non-word elements typical of these corpora. While we have no way of checking these issues manually, considering the large amounts of text, we do want to acknowledge the limits of the annotations used, something that is otherwise rarely done in corpus studies of Swedish social media (but see Brandtler 2019 for an exception).

4. Contrastive analysis of linguistic features

We explore a selection of linguistic features on the level of individual words, word classes, and sentences to empirically assess the differences and similarities between our corpora.

4.1 Personal pronouns

We begin by investigating the use of personal pronouns in our corpora. We searched for the frequent personal pronouns singular *jag* ‘I’, *du* ‘you’, *han* ‘he’, and *hon* ‘she’ in our seven corpora on the basis of their word form (case independent). Figure 1 presents their frequency per thousand words (tokens) so that we can compare their distribution among corpora of different sizes (see Table 2).

The most frequent pronoun in our social media corpora is the first-person pronoun *jag* ‘I’. The dominance of first-person pronouns is in line with Yates (1996), Radić-Bojanić (2006), and Tagliamonte & Denis (2008) for computer-mediated communication in English. It suggests that our social media corpora have a personal and involved style. This is illustrated in examples (1) and (2). Novels also have a relatively high number of first-person pronouns. This can be related to the first-person narration common in novels, or the rendering of direct speech, as in (3). News and especially Wikipedia make little use of *jag* ‘I’ – and pronouns in general – which relates to the detached style of these corpora, where writers do not refer to themselves but rather distance themselves from their utterances.

- (1) I kväll fick **jag** panik för **jag** visste inte vad **jag** skulle göra. (Blog mix: Blog mix 1998)
‘Tonight, I panicked because I didn’t know what to do.’
- (2) **Jag** har egentligen aldrig känt [sic] att den biologiska biten är viktigast däremot känner **jag** en stor sorg i att **jag** så gärna skulle vilja följa mitt barn från första början. (Familjeliv: Adoption)
‘I have never really felt that the biological part is most important, but I feel a great sadness about the fact that I would so much like to follow my child from the beginning.’
- (3) – Förlåt om **jag** är ofin, men är fröken verkligen myndig? (Novels: Norstedts)
‘– Excuse me for being rude, but is Miss really of age?’

The second-person pronoun *du* ‘you’ is most frequent in Familjeliv, Flashback, and Twitter. This might be due to the interactional nature of these platforms, where speakers can react to earlier messages, as illustrated in (4) and (5). Note that the @ sign in (5) specifies the user the message is directed to (anonymized here as @user to

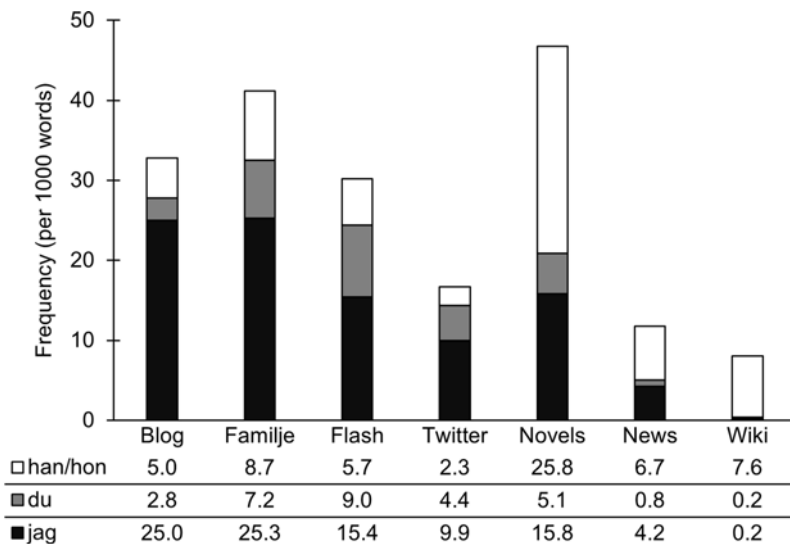


Figure 1. Personal pronouns.

protect privacy). Novels also make frequent use of second-person pronouns, particularly in passages with dialogue between characters, as in (6). News and especially Wikipedia have a very low number of second-person pronouns, which again may be related to their detached style, where writers report on facts without personal involvement.

- (4) Vad **du** behöver göra är att plocka ut alla IC-kretsar och mala dem.
(Flashback: Computer & IT)
'What you need to do is pick out all the IC circuits and grind them.'
- (5) @user Vet **du** ngt om vinet Roccolo Grassi valpolicella 2005, finns inte på bolaget.
(Twitter: Twitter mix)
'Do you know anything about the wine Roccolo Grassi valpolicella 2005, not available at Systembolaget.'
- (6) "**Du** borde vara lite mer försiktig med vad **du** säger, Ed." (Novels: Norstedts)
'"You should be a bit more careful about what you say, Ed."'

The third-person pronouns *han* 'he' and *hon* 'she' are most frequent in Novels. This might be related to the third-person narration of certain novels, as illustrated in (7). News and Wikipedia also use a relatively high proportion of third-person pronouns, illustrated in (8) and (9). This could be due to the tendency of these media to report about people.

- (7) Hennes blick vek inte undan men **hon** rodnade. (Novels: Bonniers I)
'Her eyes did not waver but she blushed.'
- (8) **Han** tog juris kandidatexamen 1948 och blev attaché vid Utrikesdepartementet (UD) 1948. (Wikipedia)

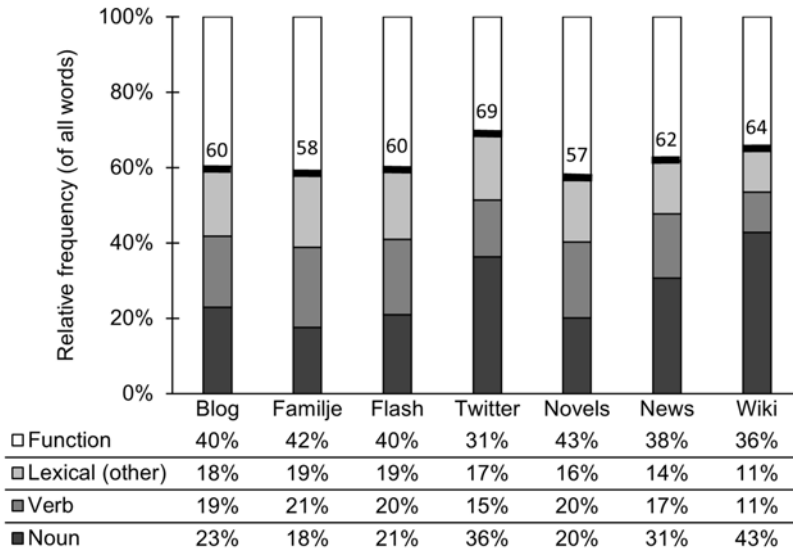


Figure 2. Nominal style and lexical density.

'He graduated with a Bachelor of Law degree in 1948 and became an attaché at the Ministry of Foreign Affairs (UD) in 1948.'

(9) **Han** fick en plastkasse med 250 000 tusen kronor i 500-lappar.

(News: Dagens Arena)

'He received a plastic bag with 250,000 thousand crowns in 500 notes.'

4.2 Nominal style and lexical density

We now move on to exploring nominal style and lexical density. Figure 2 presents the frequency of nouns (including proper nouns), verbs, other lexical words (adjectives, adverbs, participles), and function words (conjunction, determiner, infinitive marker, interjection, interrogative word, number, particle, possessive, preposition, pronoun, subjunction) based on the automatic part-of-speech annotation in the corpora. Foreign words and punctuation marks are excluded from the graph. The lexical density of each corpus is indicated by the thicker line in the bar.

All corpora have a rather similar lexical density, ranging from 57% for Novels to 69% for Twitter. This suggests that social media have a very similar ratio of lexical words and function words to other written text types. Twitter stands out with its relatively high lexical density, which could be the result of the length restriction on messages, motivating language users to prioritize content-rich lexical words over function words. If we look into the relative proportion of verbs and nouns, we find that nouns are more frequent than verbs in Twitter, News, and Wikipedia, pointing to an overall nominal style in these corpora, whereas the

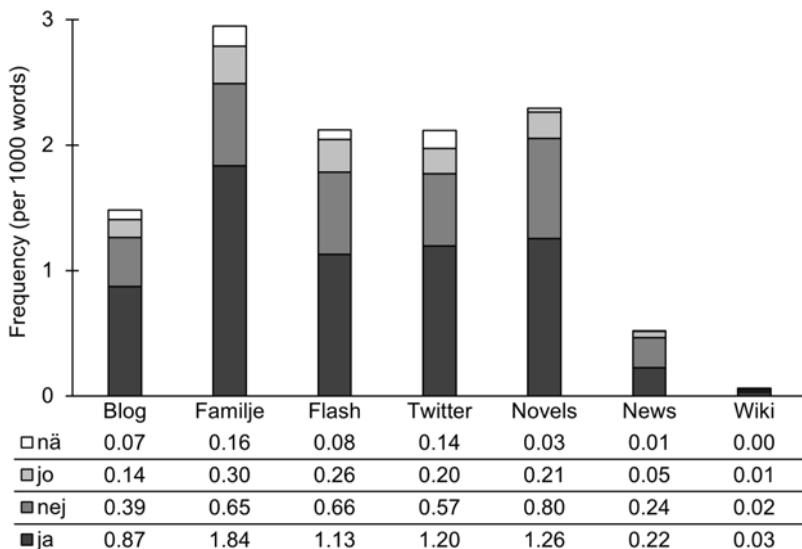


Figure 3. Answer interjections.

proportion of verbs and nouns is more balanced in Blog, Familjeliv, Flashback, and Novels.

4.3 Interjections

We proceed to investigate the use of interjections in our corpora. We first present the use of the answer interjections *ja* ‘yes’, *jo* ‘yes’, *nej* ‘no’, and *nä* ‘no’ in Figure 3.

Answer interjections are most frequent in Familjeliv, Flashback, and Twitter – the social media in our study that are designed for interaction. We illustrate part of such an interaction in (10). Novels also have a high number of answer interjections, presumably as part of direct speech, as illustrated in (11) and (12). Answer words are infrequent in News and especially in Wikipedia.

- (10) User a: Har jag fel eller? 😐
 User b: **Nej** du har rätt, att bära fönsterglas i sina bågar är inget nytt.
 (Flashback: Lifestyle)
 ‘Am I wrong or what? [neutral face emoji]’
 ‘No, you are right, wearing window glass in your frames is nothing new.’
- (11) – **Ja**, varför inte, sa Benjamin och såg på Hanna. (Novels: Norstedts)
 ‘– Yes, why not, said Benjamin, looking at Hanna.’
- (12) G. skrattar åt hennes misstänksamhet: **Jo** då, mamma har gjort en hel gryta kåldolmar för helgen. (Novels: Nordstedts)
 ‘G. laughs at her suspicion: Oh yes, mom has made a whole pot of cabbage rolls for the weekend.’

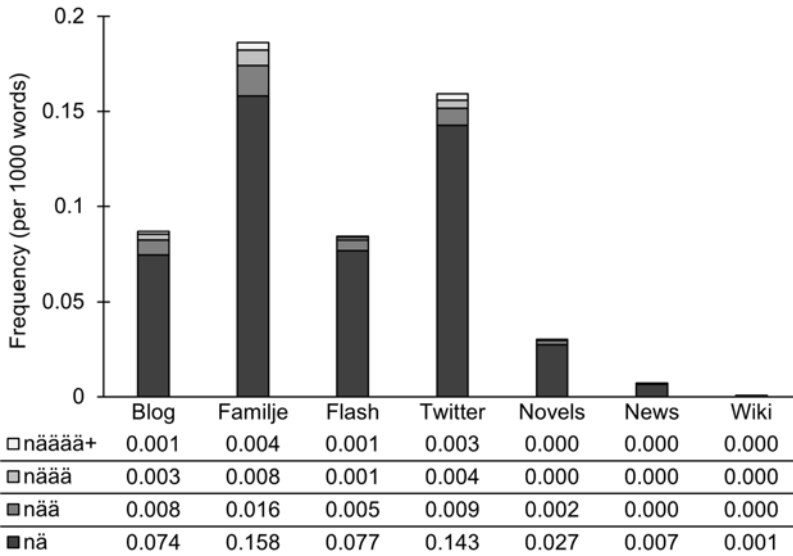


Figure 4. Answer interjection *nä* and spelling variants.

The informal spelling variant *nä* occurs predominantly in Familjeliv and Twitter. The form *nä* also occurs with a reduplication of the vowel in our social media data. Figure 4 shows the different lengths of *nä* for our corpora, given in three decimal places, given the low frequency of reduplications.

The variant *nää* is the standard spelling for emphatic, lengthened *nä*, found in all corpora to some extent, except for News and Wikipedia. This can be seen as an indication of informal language (Teleman et al. 1999:754). The longer non-standard variants with three or more vowels are only found in our social media corpora. Ortmann & Dipper (2019) see such a repetition of characters as a clear sign of conceptual orality. Example (13) and (14) show the informal, spoken-like quality of the messages with *näää* and longer variants.

- (13) Det KAN vara snyggt på vissa killar att ha i ögonbrynet, men **näää**...!
(Flashback: Lifestyle)
'It CAN be nice on some guys to have in the eyebrow, but noooo...!'
- (14) @user **nääääääääää** ... inte så. #rodnar (Twitter: Swedish Twitter 2017)
'@user noooooooooo ... not like that. #blushes'

We now turn to emotional interjections, as presented in Figure 5, both the stylistically neutral interjections *oj* 'oh, oops' and *usch* 'ugh, ew', and the profanities *helvete* 'damn', *jävlar* 'damn', and *herregud* 'oh my god', with a lower stylistic value.

Emotional interjections are most frequent in our social media corpora and in our corpus of novels, which points to an overall involved and informal style. News and Wikipedia, conversely, show very little use of these interjections. The stylistically neutral interjections *oj* and *usch* dominate in Familjeliv. As example (15) illustrates, *usch* expresses negative feelings, here in solidarity with the interlocutor. This specific

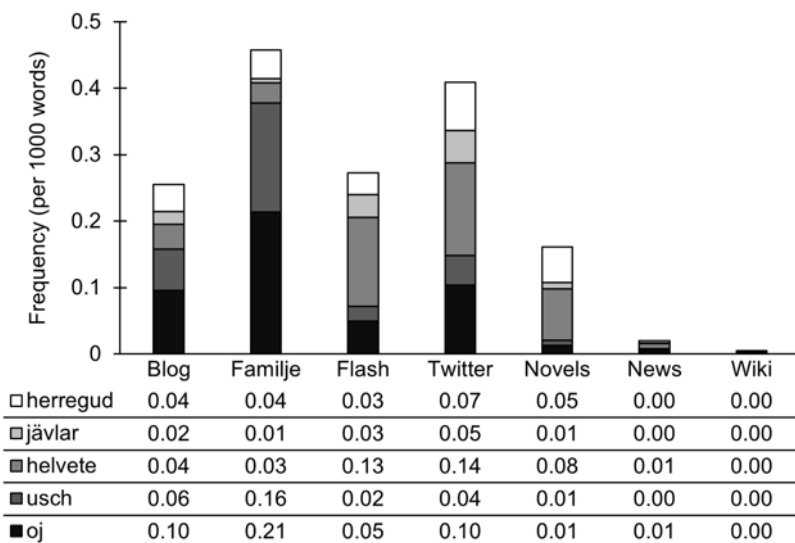


Figure 5. Emotional interjections.

use contributes to managing social relations between forum users. The profanities *jävlar*, *herregud*, and especially *helvete* dominate on Flashback and Twitter, suggesting a rougher tone of the discussion, as illustrated in example (16). The blogs have a more balanced use of both stylistically neutral and stronger interjections. Novels have a relatively high use of profanities, possibly to help dialogues come alive, as in (17).

- (15) Ta hand om dig <3 På nåt jäkla sätt klarar man ju allt på sikt <3 Men **usch**, tänker på dig. (Familjeliv: Family Planning)
 ‘Take care of yourself [heart emoticon] In some damn way you manage everything in the long run [heart emoticon] But ugh, thinking of you.’
- (16) Det är ju för **helvete** ingen raketforskning, vad tror du? (Flashback: Economy)
 ‘It’s dammit not rocket science, what do you think?’
- (17) “Javisst ja, **herregud**, vi har blivit distraherade –” Han såg på klockan på spiselhyllan. (Novels: Bonniers II)
 “‘Oh yes, my God, we’ve been distracted –” He looked at the clock on the mantelpiece.’

4.4 Sentence punctuation

We now look into the relative frequency of the period, question mark, and exclamation mark used as sentence markers in Figure 6. In some of our corpora, repeated punctuation marks are common. We only report numbers for repeated periods, as the tokenization process only correctly identifies repeated periods as one token, but misses repeated question and exclamation marks, treating them as separate tokens.

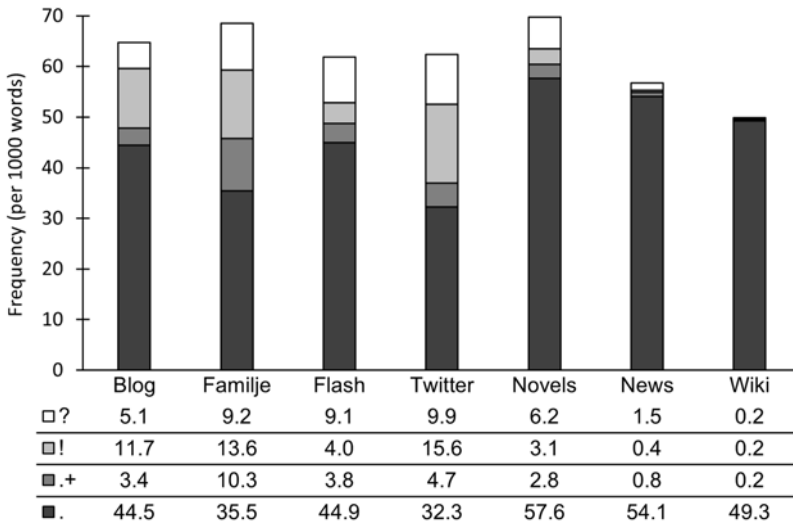


Figure 6. Sentence punctuation.

The period, marking the end of a declarative sentence, is the most frequent punctuation mark of those examined in all corpora.⁶ It has a relative frequency of around 50 instances per thousand words in most corpora, except for Familjeliv and Twitter, which only exhibit around 35 instances per thousand words. This could be the result of non-standard alternatives used for marking the ends of declarative sentences. Example (18) illustrates the non-standard use of a comma to separate two declarative sentences (known as a ‘comma splice’) and the absence of punctuation marking in the final declarative sentence of the post.

- (18) Min kompis dotter rymde från sitt familjehem i ungefär samma ålder, soc gav honom chansen att själv hämta henne (hon var hos kompisar) eller så vart det polisen (Familjeliv: Delicate Room)
 ‘My friend’s daughter ran away from her family home at about the same age, the social services gave him the chance to pick her up himself (she was with friends) or the police would.’

Question marks and exclamation marks are found predominantly in our corpora of social media and novels. This suggests that these corpora have a varied use of punctuation marks, and consequently sentence types, in contrast to News and Wikipedia. Repeated periods are also limited to our corpora of social media and novels. The most frequent repetition is that of three periods, corresponding to the standard punctuation mark known as ‘ellipsis’, illustrated in (19). Social media corpora also show longer strings of periods, sometimes up to dozens of them in a row, which is non-standard usage. Example (20) illustrates a string of 13 periods used to mark a topic shift in a post on Familjeliv.

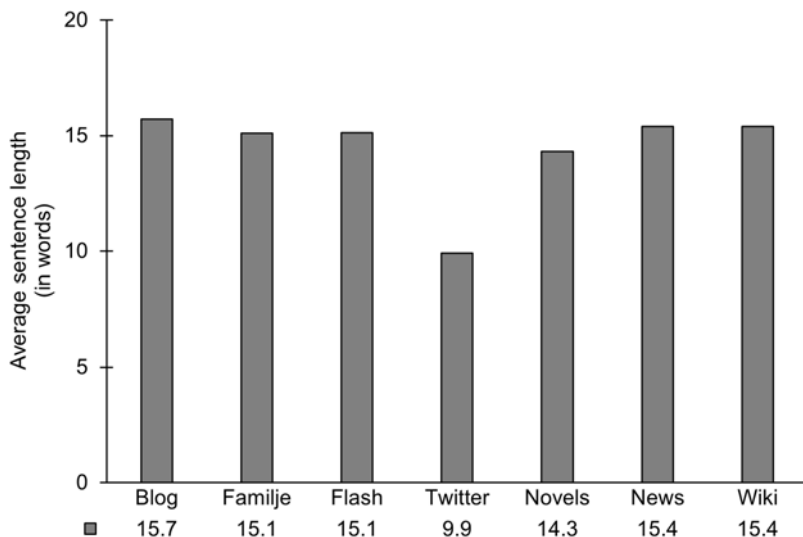


Figure 7. Average sentence length.

- (19) Moravia hade aldrig hört talas om vare sig Olle eller Sigge . . .
 (Novels: Bonniers II)
 ‘Moravia had never heard of either Olle or Sigge . . .’
- (20) Och en fråga till Finns det någon stor barnvagns-butik i
 Sverige som har ett stort sortiment med både färger och modeller?
 (Familjeliv: Pregnant)
 ‘And one more question Is there any large stroller store in
 Sweden that has a large range of both colors and models?’

4.5 Sentence length

We move on from sentence punctuation to the length of sentences, our final linguistic feature to be explored. Figure 7 presents the average sentence length, i.e. the average number of words per sentence, in our seven corpora. This metric relies on the automatic identification of sentence boundaries, based on sentence-final punctuation or line breaks (such as the end of a paragraph or message).

Figure 7 shows that the average sentence length for most of our corpora is very similar (between 14.3 and 15.7 words per sentence). Twitter is the odd one out with only 9.9 words per sentence. The tendency toward shorter sentences on Twitter most probably relates to the restriction on the number of characters per message. On the whole, average sentence length does not contribute to differentiating our corpora from each other (as opposed to earlier research by Berdicevskis 2013 and Ortmann & Dipper 2019). We therefore explore the distribution of sentences of varying length in our corpora. Figure 8 plots the frequency of sentences ranging in length from one word to 50 words.

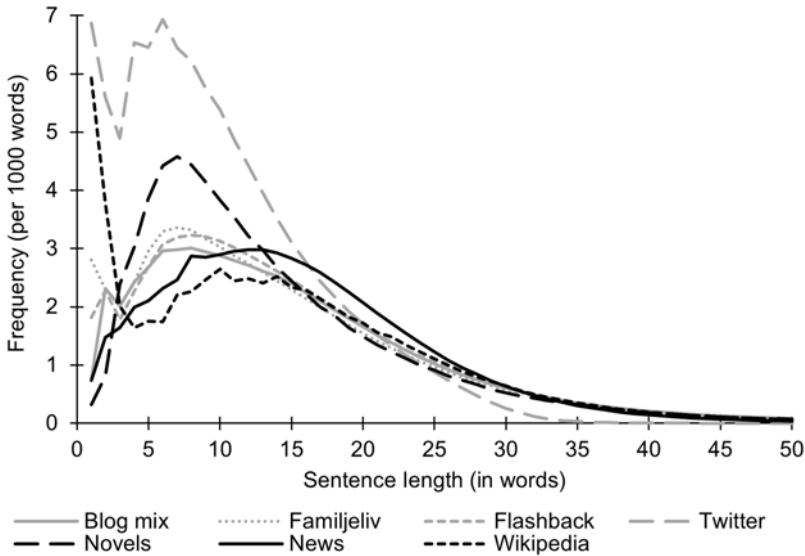


Figure 8. Distribution of sentence length.

Figure 8 shows that the distribution of sentence length follows a similar pattern of rise and fall in all of our corpora if we disregard the very short lengths of one or two. The corpora differ, however, in the sentence length at which the highest frequency is reached. Our social media corpora (gray lines) show a frequency peak of around 6 to 8 words, as does the Novels corpus. The News and Wikipedia corpora peak at around 10 to 12 words. The corpora also differ in the range of sentence lengths found. Twitter stands out with the high number of short sentences ranging between 2 and 15 words. Longer sentences are increasingly infrequent, and we hardly find any sentences longer than 35 words. Again, this is probably a direct consequence of the character restraint on the platform. Novels also have a relatively high number of short sentences. This is something of a surprise, as literary language is typically thought to be intricate and complex. As opposed to Twitter, however, we do find long winding sentences in novels. The other corpora have a more even distribution of shorter and longer sentences. Regarding sentence lengths of one or two words, these are particularly frequent on Twitter, Wikipedia, and to a lesser extent Familjeliv. The ultra-short ‘sentences’ on Twitter and Familjeliv relate to the interactive nature of these social media, where short reactions to previous messages are common. On Familjeliv, we frequently find short responses in the form of a smiley, an exclamation mark, or an interjection. On Twitter, we also find short ‘sentences’ consisting only of a link, a hashtag, or a user call with the @ sign. The one- or two-word sentences in Wikipedia turn out to be headings, which are used in this corpus to summarize and structure the information-dense content of the pages into more manageable pieces.

5. Conclusion

We set out to explore the language of social media by taking a contrastive corpus linguistic approach. More specifically, we investigated a selection of linguistic features in four corpora of Swedish social media and three other large corpora of written Swedish, all available at Språkbanken Text.

Our analysis of linguistic features revealed that social media display traits of involved, informal, non-standard, and interactional language. All our corpora of social media made frequent use of first-person pronouns and emotional interjections, indicative of involved language. The emotional interjections investigated are also informal in nature, especially the profanities that were covered in the study, which have a particularly low stylistic value. We found evidence of non-standard language in the excessive repetition of letters and punctuation, and, more specifically, in the spellings of *nä* with three or more vowels and in strings of periods longer than three. The frequent use of second-person pronouns and answer interjections points towards the interactional nature of social media.

Although all corpora of social media show the above features, they differ in their relative use of them. Familjeliv stands out with its frequent use of answer interjections, the neutral emotional interjections *oj* and *usch*, as well as the non-standard spelling of *nä* with three or more vowels, and the use of strings of periods longer than three. Twitter, in turn, makes more use of profanities, and of short replies consisting of only one or two words. A relatively high use of profanities is also typical of Flashback. These small differences in frequency across the corpora of social media indicate that they all have their own stylistic preferences.

The linguistic features discussed are not exclusive to social media but also turn out to be frequent in novels. More specifically, the novels show the same traits of involved, informal, and interactional language as social media. This is surprising as many of the studies listed in Table 1 study social media in addition, and especially in contrast, to novels. Our results indicate that the language of social media and novels is more similar in some respects than one perhaps would imagine. This said, the language of novels does differ from social media, specifically in the relative absence of non-standard language.

News and Wikipedia, in contrast, do not show much evidence of involved, informal, and interactional language. These two corpora do not often make use of first- and second-person pronouns but rather prefer detached third-person pronouns. They refrain from using answer and emotional interjections and non-standard spellings, such as repeated letters or punctuation. All of these traits point to a detached, formal, standard, and non-interactional language. This is especially the case for Wikipedia. This suggests that Wikipedia, more than News, has a large potential for serving as a reference corpus of detached, fact-based prose in Swedish.

News and Wikipedia also score high on linguistic features that indicate complexity. These corpora use nouns more than verbs, pointing to a nominal style, and have a preference for long sentences. Interestingly, Twitter scores high on some of these features: the corpus shows a nominal style and has a markedly high lexical density. These features were explained by the length restriction on Twitter posts, giving rise to an informationally dense style. The character restraint may also lie behind the low average sentence length on Twitter and its preference for short sentences. These

features make Twitter stand out as a rather hybrid genre, which shows traits of involvement, like the other social media corpora, but also signs of complex language, more like News and Wikipedia.

All in all, our analysis of linguistic features confirms previous corpus studies by empirically demonstrating the informal and non-standard nature of social media language use. Our results also nuance previous research, by revealing that traits of informal and non-standard language are distributed unevenly across the corpora of social media investigated, and are also present in other types of written texts, such as novels. We also complemented the state of the art by revealing that social media show traits of involved and interactional language, and vary in their degree of linguistic complexity. We did not explore the innovativeness of social media, as this would have required a diachronic perspective on the corpus data. This is something we would like to pursue in future research, along the lines of the pilot study in Adesam, Berdicevskis & Coussé (2024).

Acknowledgements. This work has been carried out within the Cassandra project (funded by the Marcus and Amalia Wallenberg Foundation, donation letter 2020.0060) and supported by the Swedish national research infrastructure Nationella språkbanken, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions. We thank three anonymous reviewers for their constructive comments.

Notes

1 <https://spraakbanken.gu.se/>

2 The sample combines the corpus studies cited in Wiktorsson (2018:369–370) with studies collected through our own survey in the spring of 2024. We first conducted a broad search on Google Scholar using the queries ‘Bloggmix corpus’, ‘Bloggmix korpus’, ‘Familjeliv corpus’, ‘Familjeliv korpus’, ‘Flashback corpus’, ‘Flashback korpus’, ‘Twitter corpus’, and ‘Twitter korpus’. The first ten result pages were inspected for relevant Swedish corpus studies. Undergraduate student theses were discarded. We then performed a more specific search for relevant corpus studies in the digital editions of *Språk och Stil* 2010–2022 and *Svenskans Beskrivning* 31–37 by using the terms ‘Bloggmix’, ‘Blogg’, ‘Familjeliv’, ‘Flashback’, ‘Dikussionsforum’, and ‘Twitter’.

3 We refer to the corpora of Språkbanken with a capital letter.

4 <https://ws.spraakbanken.gu.se/ws/korp/v8/>

5 <https://spraakbanken.gu.se/en/resources/corpus>

6 According to the statistics files, it is actually the most frequent punctuation mark in all corpora, except for Twitter. There, the @ sign is more frequent than the period (50.3 times per 1000 words, versus 35.3), which is a Twitter-specific sign for calling out usernames in a post. Other frequent Twitter-specific signs are the hashtag #, which is used to index keywords or topics on Twitter (14.2 times per 1000 words).

References

- Adesam, Yvonne & Aleksandrs Berdicevskis. 2021. Part-of-speech tagging of Swedish texts in the neural era. In Simon Dobnik & Lilja Øvrelid (eds.), *Proceedings of the 23rd Nordic Conference on Computational Linguistics*, 200–209. Reykjavik: Linköping University Electronic Press.
- Adesam, Yvonne, Aleksandrs Berdicevskis & Evie Coussé. 2024. Språkförändring på bar gärning: En mikrodiakron korpusstudie av pågående förändringar i stavning, lexikon och grammatik [Language change in the act: A microdiachronic corpus study of ongoing changes in spelling, lexicon and grammar]. *Svenskans Beskrivning* 38. 234–251.
- Ahlberg, Malin, Peter Andersson, Markus Forsberg & Nina Tahmasebi. 2015. A case study on supervised classification of Swedish pseudo-coordination. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, 11–19. Vilnius: Linköping University Electronic Press.

- Aichner, Thomas, Matthias Grünfelder, Oswin Maurer & Deni Jegeni. 2021. Twenty-five years of social media: A review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, Behavior, and Social Networking* 24(4). 215–222.
- Åkerblom, Sandy. 2016. Några (nya) ledtrådar till pannkaksmeningsmysteriet [Some (new) clues to the pancake sentence mystery]. *Svenskans Beskrivning* 34. 471–484.
- Allén, Sture. 1970. *Nusvensk frekvensordbok baserad på tidningstext* [Contemporary Swedish frequency dictionary based on newspaper text]. Stockholm: Almqvist & Wiksell.
- Allwood, Jens. 1998. Some frequency based differences between spoken and written Swedish. In *Proceedings of the 16th Scandinavian Conference of Linguistics*, 18–29. Turku: Turku University.
- Ameka, Felix. 1992. Interjections: The universal yet neglected part of speech. *Journal of Pragmatics* 18(2–3). 101–118.
- Berber Sardinha, Tony. 2014. 25 years later: Comparing internet and pre-internet registers. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber*, 81–107. Amsterdam: John Benjamins.
- Berber Sardinha, Tony. 2018. Dimensions of variation across internet registers. *International Journal of Corpus Linguistics* 23(2). 125–157.
- Berber Sardinha, Tony. 2022. Corpus linguistics and the study of social media: A case study using multi-dimensional analysis. In Anne O’Keeffe & Michael J. McCarthy (eds.), *The Routledge handbook of corpus linguistics*, 2nd edn, 656–674. London: Routledge.
- Berdicevskis, Aleksandrs. 2013. *Language change online: Linguistic innovations in Russian induced by computer-mediated communication*. University of Bergen PhD dissertation.
- Berdicevskis, Aleksandrs, Yvonne Adesam & Evie Coussé. 2022. Predicting short-term frequency changes in Swedish neologisms. In Elena Volodina, Dana Dannélls, Aleksandrs Berdicevskis, Markus Forsberg & Shafqat Virk (eds.), *LIVE and LEARN: Festschrift in Honor of Lars Borin*, 5–12. Göteborg: Göteborgs universitet.
- Berger, Mikael. 2020. *At ikke man får den på plass! Om mittfälsplacering av subjektspronomen i att-bisats i fastlandsskandinaviska språk [That you can’t get it in place! On midfield placement of subject pronouns in the att subordinate clause in mainland Scandinavian languages]*. In Marit Julien (ed.), *Tre uppsatser om variation och förändring*, 35–68. Lund: Lunds universitet.
- Biber, Douglas. 1986. Spoken and written textual dimensions in English: Resolving contradictory findings. *Language* 62(2). 384–414.
- Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, Douglas. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast* 14(1). 7–34.
- Biber, Douglas. 2019. Multi-dimensional analysis: A historical synopsis. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-dimensional analysis: Research methods and current issues*, 11–26. London: Bloomsbury Academic.
- Biber, Douglas & Jesse Egbert. 2018. *Register variation online*. Cambridge: Cambridge University Press.
- Blensenius, Kristian. 2013. En pluraktionell progressivmarkör? *Hålla på att jämfört med hålla på och* [A pluractional progressive marker? *Hålla på att* compared with *hålla på och*]. *Språk och Stil* 23. 175–204.
- Blensenius, Kristian & Lena Rogström. 2020. Att hantera grammatisk förändring i en deskriptiv ordbok [Managing grammatical change in a descriptive dictionary]. *Nordiska Studier i Lexikografi* 15. 81–90.
- Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp: The corpus infrastructure of Språkbanken. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, 474–478. Istanbul: European Language Resources Association.
- Brandtler, Johan. 2019. The question of form in the forming of questions: The meaning and use of clefted *wh*-interrogatives in Swedish. *Journal of Linguistics* 55(4). 755–794.
- Brandtler, Johan. 2020. Vi bara testade en hypotes . . . Ännu mer om preverbala adverbial i svenska [We just tested a hypothesis . . . More about preverbal adverbials in Swedish]. *Norsk Lingvistisk Tidsskrift* 38. 59–92.
- Bylin, Maria. 2016. Adjektivböjningens -a och -e: Ett seglivat variationstillstånd [Adjective inflection -a and -e: A long-standing state of variation]. *Språk och Stil* 26. 69–100.
- Caplan, Spencer & Kajsa Djärv. 2019. What usage can tell us about grammar: Embedded verb second in Scandinavian. *Glossa* 4(1). 1–37.
- Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen (ed.), *Spoken and written language: Exploring orality and literacy*, 35–54. Norwood: Ablex.

- Clarke, Isabelle. 2019. Functional linguistic variation in Twitter trolling. *International Journal of Speech, Language and the Law* 26(1). 1–28.
- Clarke, Isabelle. 2022. A multi-dimensional analysis of English tweets. *Language and Literature* 31(2). 124–149.
- Clarke, Isabelle & Jack Grieve. 2019. Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PLoS ONE* 14(9). 1–27.
- Collberg, Philippe & Anders Agebjörn. 2022. Svenskans obestämda plurala artikel *ena*: Lexikon, syntax och pragmatik [Swedish indefinite plural article *ena*: Lexicon, syntax, and pragmatics]. *Språk och Stil* 32. 104–136.
- Coussé, Evie, Yvonne Adesam, Faton Rekathati & Aleksandrs Berdicevskis. 2023. Hur används *de*, *dem* och *dom* i nutida skriftspråk? En storskalig korpusundersökning av nyheter och sociala medier [How are *de*, *dem*, and *them* used in contemporary written language? A large-scale corpus study of news and social media]. *Språk och Stil* 33. 39–70.
- Deumert, Ana. 2016. Linguistics and social media. In Keith Allan (ed.), *The Routledge handbook of linguistics*, 1st edn, 561–573. London: Routledge.
- Engdahl, Elisabet & Elizabeth Coppock. 2017. Absolut superlativ i samtida språkbruk [Absolute superlative in contemporary language]. *Språk och Stil* 27. 5–20.
- Engdahl, Elisabet & Anu Laanemets. 2015. Prepositional passives in Danish, Norwegian and Swedish: A corpus study. *Nordic Journal of Linguistics* 38(3). 285–337.
- Friginal, Eric, Oksana Waugh & Ashley Titak. 2018. Linguistic variation in Facebook and Twitter posts. In Eric Friginal (ed.), *Studies in corpus-based sociolinguistics*, 342–362, London: Routledge.
- Goulart, Larissa & Margaret Wood. 2019. Methodological synthesis of research using multi-dimensional analysis. *Journal of Research Design and Statistics in Linguistics and Communication Science* 6(2). 107–137.
- Halliday, Michael A. K. 1978. *Language as a social semiotic: The social interpretation of language and meaning*. London: Edward Arnold.
- Halliday, Michael A. K. 1985. *Spoken and written language*. Oxford: Oxford University Press.
- Herring, Susan C. 1996. *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*. Amsterdam: John Benjamins.
- Hillbom, Annika. 2015. Känsladjektiv i svenskan: Kategorier och figurativa användningar [Emotive adjectives in Swedish: Categories and figurative uses]. *Språk och Stil* 25. 127–159.
- Höder, Steffen. 2023. The Devil is in the schema: A constructional perspective on Swedish taboo-avoiding strategies. In Evie Coussé, Benjamin Lyngfelt, Julia Prentice & Steffen Höder (eds.), *Constructional approaches to Nordic languages*, 82–113. Amsterdam: John Benjamins.
- Hu, Yuheng, Kartik Talamadupula & Subbarao Kambhampati. 2013. Dude, srslly?: The surprisingly formal nature of Twitter's language. In *Proceedings of the International AAAI Conference on Web and Social Media* 7(1), 244–253.
- Jansson, Håkan. 2016. Från 'tandborstord' till 'memilord': Om nyord och deras belägg [From 'toothbrush words' to 'meme words': About new words and their evidence]. *Nordiske Studier i Leksikografi* 13. 359–369.
- Jensen, Bård Uri. 2007. Syntactic variables in pupils' writing: A comparison between handwritten and pc-written texts. In Gard B. Jensen, Øystein Heggelund, Margrete Dyvik Cardona, Stephanie Wold & Anders Didriksen (eds.), *Linguistics in the making: Selected papers from the Second Scandinavian PhD Conference in Linguistics and Philology in Bergen*, 165–184. Bergen.
- Julien, Marit & Helge Lødrup. 2013. Dobbelt passiv og beslektede konstruksjoner i skandinavisk [Double passive and related constructions in Scandinavian]. *Norsk Lingvistisk Tidsskrift* 31(2). 221–246.
- Katourgi, Alexander. 2022. Kolon: Ett informationsstrukturerande skiljetecken [Colon: An information structuring punctuation mark]. *Språk och Stil* 31(2). 133–158.
- Ledin, Per & Benjamin Lyngfelt. 2013. Olika hen-syn: Om bruket av hen i bloggar, tidningstexter och studentuppsatser [Different views of 'hen': On the use of 'hen' in blogs, newspaper texts and student essays]. *Språk och Stil* 23. 141–174.
- Liimatta, Aatu. 2019. Exploring register variation on Reddit: A multi-dimensional study of language use on a social media website. *Register Studies* 1(2). 269–295.
- Liimatta, Aatu. 2023. Register variation across text lengths: Evidence from social media. *International Journal of Corpus Linguistics* 28(2). 202–231.
- Ling, Rich & Naomi S. Baron. 2007. Text messaging and IM: Linguistic comparison of American college data. *Journal of Language and Social Psychology* 26(3). 291–298.

- Lomborg, Stine. 2011. Social media as communicative genres. *MedieKultur* 27(51). 55–71.
- Malm, Per. 2016. *Specificitetshypotesen: Om preteritumformer med referens till nutid och framtid* [The specificity hypothesis: About preterite forms with reference to present and future]. Göteborg: Göteborgs universitet.
- Olofsson, Joel. 2014. Argument structure constructions and syntactic productivity: The case of Swedish motion constructions. *Constructions* 9. 1–21.
- Olofsson, Joel & Julia Prentice. 2020. För tre enorma öl sedan: Befästning av semi-schematiska konstruktioner i L2-svenska [Three huge beers ago: Entrenchment of semi-schematic constructions in L2 Swedish]. *Språk och Stil* 30. 91–116.
- Ortmann, Katrin & Stefanie Dipper. 2019. Variation between different discourse types: Literate vs. oral. In *Proceedings of VarDial*, 64–79. Ann Arbor.
- Radić-Bojanić, Biljana. 2006. Fragmentation/integration and involvement/detachment in chatroom discourse. *SKASE Journal of Theoretical Linguistics* 3(1). 38–46.
- Rawoens, Gudrun. 2015. The Swedish connective *så att* 'so that'. In Andrew D.M. Smith, Graeme Trousdale & Richard Waltereit (eds.), *New directions in grammaticalization research*, 51–65. Amsterdam: John Benjamins.
- Rüdiger, Sofia & Daria Dayter. 2020. Introduction: The expanding landscape of corpus-based studies of social media language. In Sofia Rüdiger & Daria Dayter (eds.), *Studies in Corpus Linguistics*, 1–12. Amsterdam: John Benjamins.
- Skärlund, Sanna. 2014. Har *folk* blivit ett pronomen? Utvecklingen av ordet *folk* under perioden 1300–2013 [Have *people* become a pronoun? The evolution of the word *folk* in the period 1300–2013]. *Arkiv för Nordisk Filologi* 129. 245–280.
- Skärlund, Sanna. 2016. Blir *en* det nya *hen*? Om den ny(gamla) användningen av *en* som generiskt pronomen [Is *en* becoming the new *hen*? On the new (old) use of *en* as a generic pronoun]. *Svenskans Beskrivning* 34. 413–427.
- Sköldberg, Emma & Anna Helga Hannesdóttir. 2017. Svenska ord – men vilka? Om uppslagsorden i Svensk ordbok utgiven av Svenska Akademien [Swedish words – but which ones? About the entry words in the Swedish Dictionary published by the Swedish Academy]. *Svenskans Beskrivning* 35. 329–340.
- Tagliamonte, Sali A. & Derek Denis. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech* 83(1). 3–34.
- Teleman, Ulf, Staffan Hellberg, Erik Andersson & Lisa Holm. 1999. *Svenska Akademiens grammatik*, vol. 2, *Ord* [Grammar of the Swedish Academy, vol. 2, Words]. Stockholm: Svenska Akademien.
- Thyberg, Kajsa. 2020. *Det-konstruktioner i bruk: En systemisk-funktionell analys av satser med icke-referentiellt det i modern svenska* [Det constructions in use: A systemic functional analysis of clauses with non-referential *det* in modern Swedish]. University of Gothenburg PhD dissertation.
- Ure, Jean. 1971. Lexical density and register differentiation. *Applications of Linguistics* 23(7). 443–452.
- Valdeson, Fredrik. 2017. *Dativalternering i modern svenska* [Dative alternation in modern Swedish]. *Svenskans Beskrivning* 35. 355–367.
- Valdeson, Fredrik. 2021. *Ditransitives in Swedish: A usage-based study of the double object construction and semantically equivalent prepositional object constructions 1800–2016*. Stockholm University PhD dissertation.
- Vandekerckhove, Reinhild, Lisa Hilde, Darja Fišer & Walter Daelemans. 2019. Computer-mediated communication (CMC) and social media corpora: Introduction. *European Journal of Applied Linguistics* 7(2). 157–162.
- Wiktorsson, Maria. 2018. How hybrid is blog data? A comparison between speech, writing and blog data in Swedish. *Nordic Journal of Linguistics* 41(3). 367–377.
- Wiktorsson, Maria. 2022. Interjektioner eller ideofoner? En undersökning av vips och svisch [Interjections or ideophones? An examination of *vips* and *svisch*]. *Språk och Stil* 31. 72–104.
- Yates, Simeon J. 1996. Oral and written linguistic aspects of computer referencing: A corpus based study. In Susan C. Herring (ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*, 31–45. Amsterdam: John Benjamins.