SHORT COMMUNICATION

How hybrid is blog data? A comparison between speech, writing and blog data in Swedish

Maria Wiktorsson

The new forms of written online communication offer a great resource for researchers interested in language variation and use, but more large-scale systematic research into the nature of the data is needed. For instance, Swedish blog data is often described as more informal and spoken in nature than traditional edited written material but overall systematic comparisons are lacking. This short communication contributes systematic comparisons between blog data and spoken and written registers by comparing measures such as type/token ratios and word frequencies. Type/token ratios of blog texts are found to lie between those for interactive speech and formal edited writing, whereas the distribution of words from different frequency bands is closer to the written material. Comparison of the ten most frequent word forms indicates that blog data resembles formal edited writing from a structural perspective, but also suggests that further studies into features of personal involvement may provide additional insights.

Keywords: blog data, genre and register, hybrid language, speech and writing, Swedish, type/token ratio, word frequencies

Maria Wiktorsson, Malmö universitet, Kultur och samhälle K3, 205 06 Malmö, Sweden.
maria.wiktorsson@mau.se

1. INTRODUCTION

The many new forms of written communication on the Internet provide an enormous resource for linguists interested in language variation and use (Egbert, Biber & Davies 2015). There is, however, a need for more systematic large-scale studies of the nature of online data, since '[w]ithout a clear understanding of the linguistic variability of Internet texts we are severely limited in our ability to use this powerful resource for linguistic ... research' (Egbert et al. 2015:1817).

Different forms of written online communication (often generally subsumed under terms such as computer-mediated communication or netspeak – Crystal 2011) have often been described as a cross or hybrid between speech and writing, see for

instance Yates (1996), Crystal (2006, 2011) and Baron (2009); but see also Jensen (2014) for a critique of hybrid accounts. Likewise, the type of online data focused on in the current short communication, Swedish blog data, has been described as informal and spoken in nature compared to more traditional edited written material (e.g. Ahlberg et al. 2015, Hillbom 2015).

Whereas the accounts by e.g. Yates (1996) and Baron (2009) are based on systematic comparisons between different types of data, the descriptions of the Swedish blog data often appear more impressionistically based as no reference is made to systematic overall comparisons between blog data and other genres or registers. The current short communication attempts to address this gap by quantitatively investigating how Swedish blog data compares with spoken and written registers in terms of type/token ratio, ratios of words in different frequency bands, and most common words. The purpose is to encourage further, preferably more qualitative, studies of the nature of blog data.

In linguistic variation studies, the terms genre and register are not always clearly distinguished and the choice 'comes down to personal preference or tradition' (Egbert et al. 2015:1818). Biber & Conrad (2009) however make a clear distinction between genre and register. Registers are defined on the basis of situational characteristics (e.g. participants, interactivity, communicative purpose, etc.) and are analysed in terms of pervasive linguistic features functionally connected to the situational context. Genres, on the other hand, are text varieties defined by their conventional structures and features. In work on English blogs, Myers (2010a) appears to make a similar differentiation when he defines blogs as a genre (in part based on overall structuring features such as the use of links, reverse chronological order of blog entries, a column layout, etc.) while also acknowledging that the blog genre is not 'defined by a particularly [sic.?] linguistic register' (Myers 2010a:19). In work on register classification of web texts, Egbert et al. (2015) categorize blogs into different registers depending on their type, for instance news blogs and travel blogs are in the narrative register whereas opinion blogs and religious blogs are in the opinion register. Further, Biber, Egbert & Davies (2015) claim that especially personal blogs often combine different purposes, such as narration and opinion, and therefore suggest that new hybrid registers are emerging. It is clear that the blog genre displays great variation, and that more work is needed to understand how different register features function in relation to the varying purposes of texts in this genre. The current work will only scratch the surface of this material but hopes to encourage further study in doing so.

The blog data under investigation here comes from the Swedish Blog mix collection, available through the Korp interface (Borin, Forsberg & Roxendal 2012) of the Swedish Language Bank (https://spraakbanken.gu.se/eng/). On 9 April 2018, Blog mix contained Swedish blogs from 1998 to 2017 (totalling 615,658,549 tokens). Data from this collection will be compared with interactive spoken and formal written

registers, using figures for relevant features presented by Allwood (1998). Before presenting the quantitative comparison, I will review how current studies using Swedish blog data describe the material.

2. STUDIES USING SWEDISH BLOG DATA

Over the past years, a number of studies have made use of data from the Swedish Blog mix collection. The list of topics investigated include Swedish motion constructions (Olofsson 2014), passive constructions in Scandinavian languages (Julien & Lødrup 2013, Engdahl & Laanemets 2016), 'touch' adjectives in Swedish (Hillbom 2015), pragmaticalization of Swedish connective *så att* 'so that' (Rawoens 2015), current usage of the gender-neutral third-person pronoun *hen* in Swedish (Ledin & Lyngfelt 2013) and vocabulary of football commentaries (Bergh & Ohlander 2012); Swedish lexicography work (Sköldberg & Hannesdóttir 2016), supervised classification and semi-automatic discovery of Swedish pseudo-coordination constructions (Ahlberg et al. 2015), and the creation of a reference dataset for natural language processing (Eide, Tahmasebi & Borin 2016) have also been carried out using the Swedish Blog mix resource.

The Swedish Blog mix studies reviewed here often comment on the nature of the Blog mix data. As these comments shed important light on assumptions about, or observations regarding, the nature of the data, they will be of specific interest here. In general, no reference is made to systematic overall comparisons between different genres or registers which suggests that the descriptions are more loosely based on general observations of the data, or on results related to the specific features under study. The informal and casual nature of the data is often mentioned, as is its readiness to embrace novel language usage (e.g. Ledin & Lyngfelt 2013, Hillbom 2015). As many of the studies investigate novel or informal features, the motivations for the choice of empirical material are often related to the study objects. Ledin & Lyngfelt (2013:148, my translation) state that 'blogs represent a trendy and casual written language variety, where neologisms like *hen* [the Swedish gender-neutral third-person pronoun] can be presumed to gain a foothold relatively fast'. Hillbom's (2015) description is similar but adds spoken influence to the list of features: 'Blog mix ... contains texts of a casual and informal nature, with spoken influences, where novel uses of the adjectives could occur' (Hillbom 2015:135, my translation). Also Malin Ahlberg and colleagues comment on the more speech-like quality of the data:

> Since blog texts are typically informal and unedited, they contain a high degree of noise, i.e. misspellings and ungrammatical language. However, since the language of blogs is TYPICALLY CLOSER TO SPOKEN LANGUAGE THAN EDITED TEXTS, and SPCs [Swedish pseudo-coordination constructions] tend to be more frequent in spoken language, they contain

many SPCs as well as new SPC-like constructions. (Ahlberg et al. 2015:14, my emphasis)

In relation to texts such as Blog mix, Sköldberg & Hannesdóttir (2016:329, my translation) characterize them as 'unedited, spoken-like texts'. Rawoens (2015:57) claims that blog data is 'a hybrid form between written and spoken language', which, as presented in the introduction, is a common way to describe written online communication. To conclude, blog texts are described as an informal, spoken-like variety of Swedish, quick to adopt new usages.

Since language features described to be more informal or more common in speech are indeed also found in the blog data by these studies, or found to be more common in the blog data than in other types of investigated material, it does appear to be the case that the material satisfies the descriptions. Blog mix does contain informal language and constructions with less clear grammatical status than more edited texts. To what extent and in what way the genre of blog texts as a whole exhibits features more associated with spoken registers than with formal edited written registers is, however, not determined. As has been found for English blog data, the variation within the material could still be quite substantial, both in terms of how well-edited the texts are (Crystal 2006:246) and in terms of what written registers they can be classified into (Biber et al. 2015).

3. COMPARISON BETWEEN BLOG MIX AND EARLIER STUDIES OF SPOKEN AND WRITTEN SWEDISH

This section will investigate how Swedish blog data compares with spoken and written Swedish in terms of word frequencies and most commonly used words, i.e. features that have previously been used in genre or register studies (e.g. Yates 1996, Allwood 1998, Biber, Conrad & Reppen 1998). Specifically, I will compare one of the Blog mix corpora, Blog mix 2001, with the Swedish spoken and written corpora investigated by Allwood (1998). The spoken corpus represents mainly interactive spoken registers such as conversations and interviews, and the written corpus contains formal edited writing from newspapers and novels, i.e. material that falls into the broadly defined informational and narrative registers, respectively (Biber & Conrad 2009).

The comparison will take the results presented by Allwood as its starting point, and corresponding figures will be determined for Blog mix 2001, by making use of the associated statistics file (https://svn.spraakdata.gu.se/sb-arkiv/pub/frekvens/stats_ BLOGGMIX2001.txt). As different principles underlie the figures obtainable from the Swedish Language Bank and Allwood's (1998) numerical results (for instance regarding word form statistics), manual re-calculations and re-groupings were necessary to arrive at comparable figures. The details of these manipulations are specified below in connection with the figures presented.

Blog mix 2001 was selected because its word count is comparable to the corpora used by Allwood (1998). Allwood's (1998) spoken corpus contains 276,391 words and his written corpus, 271,216 words. Blog mix 2001 contains 287,342 words when punctuation is removed from the original token count of 326,659.

3.1 Type/token ratios

In order to explain how and why the statistics file associated with Blog mix 2001 was manipulated we have to understand the underpinnings of the statistics presented by Allwood (1998). Allwood's token count is simply the total number of words in the corpora. Punctuation is not included in the word count for the written data. In the spoken data, the token count includes all transcribed words, including feedback words such as *ja* 'yes', *m* 'm' and own communication management words such as *ä* 'eh'.

The type counts used by Allwood (1998) are based directly on word form, at least for the written data. This means that no homonym differentiation has taken place to calculate the overall types in the written data. In this count, *springa* 'run/crack' is counted as the same type regardless of whether we are dealing with the verb 'run' or the noun 'crack'. The spoken word forms are, however, disambiguated according to the written word form in cases where the same spoken form corresponds to two (or more) written forms. For example, the spoken form /å/ corresponds to either *och* 'and' or *att* 'to' in writing. In the type count reported for speech, the spoken word forms are disambiguated according to their written counterparts.

The word frequency statistics for Blog mix 2001 is not counted on direct word forms, but rather lists, as unique lines in the file, word forms with the same Part-of-Speech tag (POS-tag[i]) and Lemgram label. The Lemgram label gives the lemma and inflectional table for each word form, and includes both unitary lemmas and multiword lemmas (when relevant). This means that the same word form, depending on how it has been analysed in terms of POS and Lemgram, will have a separate count in the file. This holds for the same word form used as an adjective or verb, such as *rädda* which can mean both 'afraid' and 'save', see Table 1. It also holds

Word form	POS	Lemgram	Count
rädda 'afraid'[a]	JJ.POS.UTR+NEU.PLU.IND+DEF.NOM	|rädd..av.1|	19
rädda 'save'[b]	VB.INF.AKT	|rädda..vb.1|	11

[a]The POS label for *rädda* 'afraid' marks the word form as an adjective (JJ) in the positive form of comparison (POS), agreeing with both the non-neuter (UTR) and neuter (NEU) genders in plural (PLU), and with both indefinite (IND) and definite (DEF) definiteness, in the nominative (NOM) case. The lemgram label identifies the lemma (and the associated inflectional table) of the adjective *rädd* 'afraid', and is formed from the lemma (rädd), the part of speech (av = adjective), and a disambiguating numeral.
[b]The POS label for *rädda* 'save' marks the word form as a verb (VB) in the infinitive (INF), active (AKT) form. The lemgram label identifies the verb (vb) lemma *rädda* 'save'.

Table 1. Word class disambiguation of *rädda* in the Blog mix 2001 statistics file.

Word form	POS	Lemgram	Count
och	KN	|och..kn.1|	8821
och	KN	|och..kn.1|och_så_vidare..abm.1|	14
och	KN	|och..kn.1|både_och..knm.1|	188
och	KN	|och..kn.1|till_och_med..abm.1|till_och_med..ppm.1|	24
och	KN	|och..kn.1|i_och_med..ppm.1|	12
och	KN	|och..kn.1|saker_och_ting..pnm.1|	10
..	. . .	. . .	. . .

Note: The POS label KN stands for conjunction. The lemgram label first gives the specific label for the lemma *och* 'and' (kn = conjunction) and, in relevant cases, indicates when *och* 'and' is part of a multiword expression. The multiword expression starts the complex lemgram label, and is followed by a specification of the part of speech of the unit (abm = multiword adverb; knm = multiword conjunction; ppm = multiword preposition; pnm = multiword pronoun).

Table 2. Complex lemma disambiguation of *och* 'and' in the Blog mix 2001 statistics file.

	Blog mix 2001	Allwood's (1998) spoken corpus	Allwood's (1998) written corpus
Types	30869	18406	39638
Tokens	287342	276391	271216
Ratio	10.7%	6.7%	14.6%

Table 3. Comparison of word form type/token ratios in Blog mix 2001, speech and writing.

for usages in different complex lemmas, such as for the word form *och* 'and', see Table 2.

To obtain type data for Blog mix 2001 comparable to Allwood (1998), the statistics file from the Swedish Language Bank was re-sorted on word form only, and the numbers were grouped together. So, in the type count, the POS and Lemgram information available was disregarded. Table 3 compares the type/token counts and ratios in the Blog mix data with those reported for speech and writing by Allwood (1998).

As type–token ratios measure lexical diversity, we can conclude that the blog data does appear to lie between the spoken and written material when it comes to how many different words are used. There are fewer types used than in writing but more than in speech. Thus, the type of written language variety represented by these blog texts appear to employ a slightly narrower range of words than more traditional writing, but it cannot be said to be closer to speech in terms of vocabulary use.

3.2 Frequency-based rank

Another measurement of relevance from Allwood's (1998) comparison of speech and writing is to what degree words in different frequency ranges are used in the corpus. This measures to what degree a text employs words of different frequencies.

Frequency-based rank	Blog mix 2001	Speech (Allwood 1998)	Writing (Allwood 1998)
1–10	20.5%	23.3%	19.4%
11–50	19.0%	28.7%	18.8%
51–100	7.8%	10.3%	6.8%
101–1000	24.0%	22.2%	20.7%
1001–10000	20.2%	12.5%	21.3%
10001–30869 (Blog mix 2001)	8.5%		
10001–18406 (Speech)		3.0%	
10001–39638 (Writing)			13.0%

Table 4. Corpus share of words from different frequency bands.

A higher degree of the most common words means that these are employed more in the texts. Table 4 compares Allwood's figures with the figures for Blog mix 2001.

In terms of frequency-based ranks of word forms only (no homonym disambiguation) the Blog mix data appears closer to writing than to speech in terms of the degree of usage of the 10, 50 and 100 most common words. The most infrequent words (above rank 10,000) make up a smaller proportion in the blog data than in the written data, but larger than in the spoken data.

The blog data appears closer to writing here, which should be compared with the overall type/token count presented above, where the blog variety occupied a position mid-way between speech and writing. We can conclude that blog texts do not repeat the most common words to the same degree as interactive speech, but rather display a closer rank distribution to the written data.

3.3 Most frequent word forms

The most frequent words in the different corpora can reveal interesting differences and shed more light on how blog data compares to the spoken and written data. Table 5 lists the 10 most frequent words in the three text types. As the total number of words in the three text collections are roughly comparable, the absolute numbers are included here for direct comparison.

Unsurprisingly, all the words in these lists are function words. We can also observe that six of the function words (*och* 'and', *det* 'it', *att* 'that, to', *är* 'is', *på* 'on' and *som* 'that, which') are shared between all three lists. Even though there are rank differences, it seems safe to assume that these six words make up some kind of common core repository of function words used in many different types of Swedish constructions, regardless of register.

There are more commonalities in terms of rank of the six common words between the blog data and writing than between blog data and speech. Three of the words – *och* 'and', *att* 'that, to', *på* 'on' – have the same rank in blogs and writing. These

Rank	Blog mix 2001	Freq.	Speech (Allwood 1998)	Freq.	Writing (Allwood 1998)	Freq.
1	och 'and'	9188	de (det) 'it'	16898	och 'and'	8378
2	det 'it'	7325	e (är) 'is'	6666	i 'in'	7683
3	att 'that, to'	7234	å (och) 'and'	6645	att 'that, to'	6638
4	i 'in'	5992	så 'so'	6424	det 'it'	5887
5	jag 'I'	5582	ja (jag) 'I'	5977	som 'that, which'	4808
6	är 'is'	5446	att 'that, to'	5579	en 'a, an, one'	4638
7	på 'on'	5165	ja 'yes'	4475	på 'on'	4161
8	som 'that, which'	5041	på 'on'	4043	är 'is'	3923
9	en 'a, an, one'	4628	som 'that, which'	3866	med 'with'	3243
10	för 'for'	3356	man 'one'	3794	av 'of, by'	3133

Table 5. The ten most frequent words in the three corpora.

are words of a type Allwood (1998:Section 3) suggests are needed to 'construct a complex phrase and sentence structure which is more typical of written than of spoken language'. Additional prepositions occur in the written variety (*med* 'with' and *av* 'of, by'), in the blog data (*för* 'for'), and in both writing and blog data (*i* 'in'). As prepositions often have a function in more complex phrase and clause structures, the higher token and type count for these would suggest a more complex structure in the blog and written data compared to the spoken data.

First-person pronouns are found among the register features associated with involved production, that is, production with 'an involved, non-informational focus, related primarily to a primary interactive or affective purpose and online production circumstances' (Biber et al. 1998:149). Face-to-face conversations are typical cases of this kind of production and rank high on involvement features (Biber et al. 1998:152). Since Allwood's (1998) spoken data is of an interactive type, the high frequency of the first-person pronoun *jag* 'I' is not surprising. Worth noting is that the blog data display similarities to the spoken data in this respect. The pronoun *jag* 'I' is found on the same rank position in Blog mix as in interactive speech. Naturally, the frequency of a single pronoun is not enough to determine the involved nature of the blog genre. Further work is needed here, perhaps on features 'that can be interpreted as reflecting interpersonal interaction and the involved expression of personal feelings and concerns' (Biber et al. 1998:150) such as first- and second-person pronouns, *wh*-questions, emphatics, etc. (Biber et al. 1998:149–150).

4. LIMITATIONS

The kinds of overall quantitative measurements compared in the current short communication cannot reveal the full nature of the blog genre. Much more work

of a qualitative nature is needed to better understand how bloggers make use of features associated with different spoken and written registers.

5. CONCLUSIONS

Based on the comparisons made above, it cannot be claimed that the blog data in the Blog mix 2001 corpus is closer to speech than to writing. In terms of overall type/token ratio, the blog data rather takes a middle position between the spoken and written registers compared with here. In terms of vocabulary variance, the blog texts do not rely on the most common words to the same extent as interactive speech does, but rather display a distribution between different frequency ranks more similar to formal edited writing. The ten most frequent words in each of the three text types indicate that the blog data is more like writing from a basic structural perspective, but perhaps display certain personal involvement features normally associated with interactive speech.

To sum up, blog texts are not spoken in nature from the perspectives investigated here, but they are not entirely like more formal and edited writing either. They display less lexical richness than formal edited writing, and, as attested by the studies cited in Section 2, contain more informal language and marginal grammar. In conclusion, more work is needed to better understand how the blog genre makes use of features from different registers. Features reflecting interpersonal interaction may be especially relevant to investigate, as the perceived spoken nature may, at least partly, result from bloggers 'choosing features appropriate to the interpersonal rhetoric of the genre, and thus using features that are also more likely to be associated with face-to-face communication' (Myers 2010b:270).

ACKNOWLEDGEMENTS

NOTE

1. The full list of POS tags used in the Swedish Language Bank corpora can be found through https://spraakbanken.gu.se/korp/markup/msdtags.html.

REFERENCES

Ahlberg, Malin, Peter Andersson, Markus Forsberg & Nina Tahmasebi. 2015. A case study on supervised classification of Swedish pseudo-coordination. *Proceedings of the 20th*

Nordic Conference of Computational Linguistics, NODALIDA 2015, 11–19. Linköping: Linköping University Electronic Press.

Allwood, Jens. 1998. Some frequency based differences between spoken and written Swedish. *Proceedings of the 16th Scandinavian Conference of Linguistics*, 18–29. Turku: Turku University, Department of Linguistics.

Baron, Naomi. 2009. The myth of impoverished signal: Dispelling the spoken language fallacy for emoticons in online communication. In Jane Vincent & Leopoldina Fortunati (eds.), *Electronic Emotion: The Mediation of Emotion via Information and Communication Technologies*, 107–135. Bern: Peter Lang.

Bergh, Gunnar & Sölve Ohlander. 2012. Free kicks, dribblers and WAGs: Exploring the language of "the people's game". *Moderna språk* 106(1), 11–46.

Biber, Douglas & Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, Douglas, Jesse Egbert & Mark Davies. 2015. Exploring the composition of the searchable web- a corpus-based taxonomy of web registers. *Corpora* 10(1), 11–45.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp: The corpus infrastructure of Språkbanken. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC 2012), 474–478. Istanbul: ELRA

Crystal, David. 2006. *Language and the Internet*, 2nd edn. Cambridge: Cambridge University Press.

Crystal, David. 2011. *Internet Linguistics: A Student Guide*. London & New York: Routledge.

Egbert, Jesse, Douglas Biber & Mark Davies. 2015. Developing a bottom–up, user-based method of web register classification. *Journal of the Association for Information Science and Technology* 66(9), 1817–1831.

Eide, Stian R., Nina Tahmasebi & Lars Borin. 2016. The Swedish Culturomics Gigaword Corpus: A one-billion-word Swedish reference dataset for NLP. *Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts, Proceedings of the Workshop*, 8–12. Linköping: Linköping University Electronic Press.

Engdahl, Elisabet & Anu Laanemets. 2016. Opersonlig passiv i danska, norska och svenska–en korpusstudie [Impersonal passive in Danish, Norwegian and Swedish: A corpus study]. *Norsk Lingvistisk Tidsskrift* 33(2), 129–156.

Hillbom, Annika. 2015. Känseladjektiv i svenskan: kategorier och figurativa användningar [Touch adjectives in Swedish: Categories and figurative uses]. *Språk och stil* 25, 127–159.

Jensen, Eva Skafte. 2014. Tale er tale; skrift er skrift. Om skriftsproget i de nye medier [Speech is speech; writing is writing: About the written language of new media]. *Nydanske Sprogstudier* 46, 11–38.

Julien, Marit & Helge Lødrup. 2013. Dobbel passiv og beslektede konstruksjoner i skandinavisk [Double passive and related constructions in Scandinavian]. *Norsk lingvistisk tidsskrift* 31(2), 221–246.

Ledin, Per & Benjamin Lyngfelt. 2013. Olika hen-syn. Om bruket av hen i bloggar, tidningstexter och studentuppsatser [Different *hen*-view: About the use of *hen* in blogs, newspaper texts and student essays]. *Språk och stil* 23, 141–174.

Myers, Greg. 2010a. *The Discourse of Blogs and Wikis*. London: Continuum.

Myers, Greg. 2010b. Stance-taking and public discussion in blogs. *Critical Discourse Studies* 7(4), 263–275.

Olofsson, Joel. 2014. Argument structure constructions and syntactic productivity: The case of Swedish motion constructions. *Constructions* 1(7), 1–17.

Rawoens, Gudrun. 2015. The Swedish connective *så att* 'so that'. *New Directions in Grammaticalization Research* 166, 51–65.

Sköldberg, Emma & Anna Helga Hannesdóttir. 2016. Svenska ord – men vilka? Om uppslagsorden i Svensk ordbok utgiven av Svenska Akademien [Swedish words: But which? About the entry words in the Swedish Academy dictionary]. *Svenskans beskrivning 35, Förhandlingar vid trettiofemte sammankomsten*, 329–340. Göteborg: Göteborgs universitet.

Yates, Simeon. 1996. Oral and written aspects of computer conferencing. In Susan Herring (ed.), *Computer-mediated Communication: Linguistic, Social, and Cross-cultural Perspectives*, 22–46. Amsterdam: John Benjamins.