Investigating Interaction Signs across Genres, Modes and Languages: The Example of OKAY

Laura Herzberg, Angelika Storrer
Department of German Linguistics, University of Mannheim
Schloss, Ehrenhof West, D-68131 Mannheim
E-mail: lherzber@mail.uni-mannheim.de, astorrer@mail.uni-mannheim.de

Abstract

This paper presents results of a case study that compared the usage of OKAY across genre types (Wikipedia articles vs. talk pages), across modes (spoken vs. written language), and across languages (German vs. French CMC data from Wikipedia talk pages). The cross-genre study builds on the results of Herzberg (2016), who compared the usage of OKAY in German Wikipedia articles with its usage in Wikipedia talk pages. These results also form the basis for comparing the CMC genre of Wikipedia talk pages with occurrences of OKAY in the German spoken language corpus FOLK. Finally, we compared the results on the usage of OKAY in German Wikipedia talk pages with the usage of OKAY in French Wikipedia talk pages. With our case study, we want to demonstrate that it is worthwhile to investigate interaction signs across genres and languages, and to compare the usage in written CMC with the usage in spoken interaction.

Keywords: interaction signs, cross-lingual CMC study, Wikipedia talk pages

1. Background and Motivation

Interaction signs are elements that are not integrated in the syntactic structure of utterances, but serve as devices for discourse management: they can be used to express reactions to a partner's utterances or to display emotions. The category "interaction sign" was defined in Beißwenger et al. (2012), building on the grammar framework of the "Grammatik der deutschen Sprache" (henceforth *GDS*), which already included interjections ("hm", "well", "oh my god", "oops") and responsives ("yes", "no", "okay"). This framework was expanded with categories which have similar functions as interjections and responsives but typically occur in computer-mediated communication (CMC), e.g. emoticons, addressing terms (@USERNAME), action words ("lol", "grin") etc. (cf. Beißwenger et al., 2012).

The focus of this paper is on OKAY, which is an interesting object of study because it is used in many languages with a wide range of functions (cf. Figure 2). OKAY is not a CMC-specific interaction sign (like emoticons or "lol"), but is used in both written and spoken language. In our studies, the meta-lemma OKAY represents the different variants of spelling and pronunciation. Using OKAY as an example, we want to demonstrate that comparing the usage of interaction signs in speech corpora with its usage in written CMC corpora can yield interesting results. In our cross-genre and cross-lingual studies, we also explore which spelling variants are preferred by the users and whether these variants are compliant with spelling rules.

Most of the previous work on OKAY deals with spoken language: Schegloff/Sacks (1973) investigate OKAY in pre-closing sequences of spoken conversation. The studies of Beach (1993) and Bangerter et al. (2003) examine the usage of OKAY in phone calls. Levin/Gray (1983) describe the usage of OKAY in lecturer's presentations. Condon/Čech (2007) investigate the role of OKAY in decision making processes, comparing face-to-face interaction with CMC data. All these studies deal with the usage of OKAY in English. Studies on other languages are rare, although OKAY is used in many languages: Delahaie (2009) studies the usage of OKAY as an agreement marker in the learning of French as a foreign language. Kaiser (2011) investigates the usage of OKAY in German spoken doctor-patient communication. Cirko (2016) describes the usage of OKAY in German examination talks.

In our paper, we investigate the usage of OKAY across genre types (comparing CMC with text genres), across modes (comparing the usage in spoken interaction and written CMC), and across languages (comparing the same CMC genre in German and French). The cross-genre study builds on the results of Herzberg (2016), who compared the usage of OKAY in German Wikipedia article talk pages with its usage in Wikipedia articles. These results also form the basis for contrasting the usage of OKAY in written CMC and in spoken interaction (using data from the German speech corpus FOLK). Finally, we compare the usage of OKAY in the German Wikipedia talk pages with its usage in French Wikipedia talk pages.

2. Cross-genre study

2.1 Corpus Data

For the cross-genre study we compared data from two linguistically annotated Wikipedia corpora (cf. Margaretha/Lüngen, 2014): a corpus with German Wikipedia articles (Wiki-A-de; appr. 797 million tokens) and a corpus with German Wikipedia article talk pages (Wiki-D-de; 310 million tokens). Wikipedia articles represent a text genre (monologous structure, standard language etc.), while talk pages have features of CMC genres (dialoguous structure, informal writing style with non-standard language etc., cf. Storrer, 2017). The two corpora were downloaded from the Institute for the

German Language (IDS) and queried in RAPIDMINER-KOBRA[1].

2.2 Classification: Categories and Procedure

(1) First, we analysed the frequency of different spelling variants of OKAY in both corpora. The assumption was that OKAY is quite more frequent in the CMC corpus Wiki-D-de due to the dialogical structure and conversation-like nature of Wikipedia discussions. Different spelling variants had been queried and combined to draw the samples for the article and the talk pages (cf. Herzberg, 2016 for details). Since not all spelling variants occurred equally in both corpora, the two samples differ in their totals. The procedure resulted in a Wiki-A-de sample of 6,336 OKAY occurrences in total, and in a Wiki-D-de sample of 10,554 occurrences in total. All occurrences in both samples were manually checked and the false positives were sorted out. The distribution of true and false positives is illustrated in Table 1. It shows absolute frequencies as well as normalised frequencies as *pmw* values (occurrence per million words). Three types of false positives were distinguished: a) OKAY was mentioned as a word, e.g. in an article about interjections, b) OKAY was cited, e.g. in a song title or c) spelling variants of OKAY were homographic with abbreviations of proper names, such as a volcano ("Ok [...] is a shield volcano in Iceland")[2].
(2) Second, each spelling variant had been investigated individually. The two categories "conformant vs. non-conformant" and "speedy vs. non-speedy" served as objects of study. Because CMC writing is less norm-conformant, we expected to find spelling variants that do not comply with the German spelling norm. In German *okay*, *Okay*, *o. k.* and *O. K.* are the norm-conformant spelling variants[3]. It has to be noted, that the variants "o. k." and "O. K." have to display a blank space between *O* and *K* to be norm-conformant. Therefore, the spelling variants *ok*, *OK*, *Ok*, *o.k.*, and *O.K.* are non-conformant spellings.
Another hypothesis was that CMC users prefer "speedy" spelling variants (*ok*, *Ok*, *OK*) because speed writing is a general feature of CMC. We classified *ok*, *OK* and *Ok* as "speedy" and all other variants as "non-speedy".

2.3 Results and Discussion

(1) The results of the cross-genre frequency study on OKAY are presented in Table 1.

	true positives		false positives	
	abs.	pmw	abs.	pmw
Wiki-A-de	25	0.03	6,311	7.92
Wiki-D-de	8,248	26.62	2,306	7.44

Table 1: Distribution of true and false positives of OKAY in the German Wikipedia.

As expected, OKAY is quite more frequent in the CMC corpus (talk pages) than in the text corpus (Wikipedia articles). Interestingly, the two corpora considerably differ in their number of false positives: in the CMC sample, 2,306 (21.8 %) were classified as being false positives. In the text sample 6,311 (99.6 %) occurrences of OKAY turned out to be false positives: only 25 (0.4 %) of all occurrences were true positives. As 25 items is a very small data set, we restricted our studies on the frequency of spelling variants on the CMC sample.
(2) Table 2 shows the results of the studies on norm-conformance and frequency of OKAY spelling variants in the German CMC corpus Wiki-D-de and in the French CMC corpus Wiki-D-fr. In this section, we discuss the results of the German data; the cross-lingual aspects are treated in section 4.3[4].

Spelling Variant	Norm-conformance		Frequency Wiki-D-de		Frequency Wiki-D-fr	
	DE	FR	abs.	pmw	abs.	pmw
OK			17,796	57.43	9,281	67.69
ok			16,048	51.78	5,476	39.94
Ok			15,431	49.79	7,495	54.67
okay	✓		8,421	27.17	86	0.63
Okay	✓		8,287	26.74	163	1.19
o. k.	✓		96	0.31	0	0
O. K.	✓		86	0.28	6	0.04
o.k.			80	0.26	0	0
O.K.		✓	21	0.07	3	0.02

Table 2: Frequency and norm-conformance of OKAY spelling variants in German and French.

The results in Table 2 clearly support the assumption that non-conformant variants are more frequently used than the conformant ones in the German CMC corpus. Moreover, the results support the hypothesis that the three speedy variants *ok*, *OK*, and *Ok* are preferred, although they do not conform to German spelling rules.

3. Cross-modal study

3.1 Corpus Data

There are significant differences between the usage of interaction signs in spoken and written language. In spoken interaction intonation plays a crucial role in interpreting a positive, negative, or doubting evaluation expressed by an interaction sign. Interaction signs are relevant for organizing turn-taking in spoken interaction: hearers use interaction signs to encourage the floor holder to continue (so-called "continuers", cf. Schegloff, 1982 p. 81). While these functions have been widely investigated in spoken language (see cited works in section 1), studies on CMC or cross-modal studies are still rare.

[1] Details on the queries are provided in Herzberg (2016).
[2] Cf. https://en.wikipedia.org/wiki/Ok_(volcano) [15.06.17].
[3] Cf. Duden-Rechtschreibung, 2013 p. 781.
[4] We integrated the French data in Table 2 in order to save space. The table presents absolute frequencies as well as normalised frequencies as *pmw* values.

In our cross-modal case study, we compared data from the CMC corpus Wiki-D-de (310 million tokens), with spoken interaction data taken from the German FOLK corpus (1.9 million tokens). The speech data was queried automatically via the DGD.

3.2 Classification: Categories and Procedure

(1) We distinguished between two main functional categories: OKAY as a syntactic unit (used in predicative, adverbial, attributive function or as a noun) and OKAY as an interaction sign (used in responsive, reactive, interrogative, and structural function). In Herzberg (2016), all true positives in the Wiki-D-de sample described in section 2 (8,248 occurrences in total, cf. Table 1) have been classified as follows: 5,045 (61.2 %) occurrences are used as interaction signs and 3,203 (39.8 %) as syntactic units. An interesting finding concerns the functional category "responsive", i.e. a (positive) answer to a polar question. This function is described as being the main function of OKAY in the German grammar GDS (1997) p. 63. However, the study revealed that only a very small amount (20 occurrences, i.e. 0.4 % of all interactive OKAY occurrences) in the examined Wiki-D-de data were used as responsives. We assumed that this mismatch between the Grammar description and our data was due to the fact that the classifications in this Grammar refers to the usage of OKAY in spoken interaction. We thus used data from the FOLK corpus to investigate whether the responsive function of OKAY is a main function in spoken interaction. We manually checked how often the responsive function of OKAY was used in a FOLK corpus sample with 1,500 occurrences of OKAY.
(2) In a second study we analysed the positions of OKAY in samples taken from FOLK, Wiki-D-de, and the French Wikipedia talk pages (Wiki-D-fr; 137 million tokens) with 500 occurrences in each sample. These samples only contain true positives; false positives have been manually sorted out. The data has then been classified according to four positional categories: initial (directly at the beginning of a [5]post/utterance); middle (within a post/utterance); final (end of a post/utterance) and standalone (OKAY forms post/utterance).

3.3 Results and Discussion

(1) The analysed FOLK sample had a similar outcome as the study on the CMC data: Only 15 (1 %) of the 1,500 examined occurrences are used as responsives. In both corpora, the responsive function which is claimed to be the main function in the GDS grammar description, only rarely occurs in both written CMC and spoken language. These results demonstrate that it is worthwhile to further evaluate assumptions about the functions of interaction signs on the basis of corpus data.

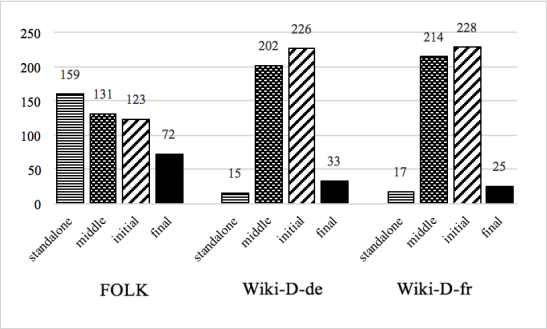

Figure 1: Positional distribution of OKAY in spoken and CMC interaction[6].

(2) The results of our comparison of positions in Figure 1 reveal significant differences between the two modes. Whereas OKAY is used variably and nearly equally often in German utterances across the three categories "standalone" (32.8 %), "middle" (27.0 %) and "initial" (25.4 %), the German CMC data presents a different picture: OKAY is preferably used at the beginning (47.5 %) and within (41.6 %) a post. These two positions make up nearly 90 % of all investigated occurrences. Interestingly, the positional distribution patterns in the German and French CMC data are quite similar[7].
There are two possible explanations for these results that have to be verified in further work: (a) The standalone position is typical for "continuers" (see above) and the final position is typical for the usage of OKAY as a tag question. Both functions are particularly relevant for organising turn-taking in spoken interaction. This may explain the lower rate of standalone and final positions in the CMC data, where turn-taking mechanism are substituted by other mechanisms of interaction management (cf. Beißwenger, 2008). (b) As it is shown in Figure 3, OKAY is mostly used as an interaction sign in the speech data from FOLK. In the two CMC corpora however, OKAY is also used as a syntactic unit. These syntactically integrated units (nouns, adverbials, predicatives) often occur in a middle position. This may be one factor to explain the higher rate of middle positions in CMC corpora in the results presented in Figure 1.
To get a clearer image of the differences in the usage of OKAY in spoken and written interaction, we want to annotate the functional and the positional categories presented in Figure 2 on two different layers and explore correlations between the positional und functional categories in more detail.

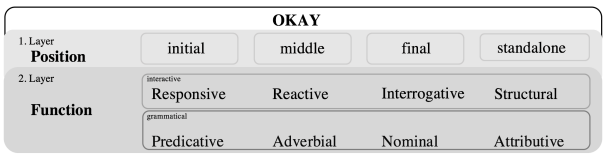

Figure 2: Formal and functional annotation categories.

[5] Following the proposals of the TEI CMC group, we use the term "posts" for units in CMC interaction (cf. Beißwenger et al., 2012). The segments in spoken interaction are termed as "utterances".

[6] The figures contain absolute frequencies of true positives.

[7] Cross-lingual aspects are treated in section 4.3.

4. Cross-lingual study

4.1 Corpus Data

In our cross-lingual study we compared the corpus of German Wikipedia talk pages (Wiki-D-de; 310 million tokens) with a corpus of French Wikipedia talk pages (Wiki-D-fr; 137 million tokens). Both corpora are available within the German Reference Corpus DeReKo at the IDS. The data has been queried automatically via COSMAS II.

4.2 Classification: Categories and Procedure

(1) In a first study, we manually classified two samples of German and French, each containing 500 OKAY occurrences, in three categories: "syntactic units", "interaction signs" (cf. 3.2) and "others". We assumed that the usage of OKAY as a syntactic unit, signalling a deeper integration of the loan word in the host language system, is less frequent in the French corpus.
(2) The focus on the second investigation was again on spelling variants. We expected that the speedy and non-conformant variants are also preferred in the French CMC corpus. Similar to the study of German, *ok*, *OK* and *Ok* were classified as speedy variants whereas *okay*, *Okay*, *o.k.*, *O.K.*, *o. k.* and *O. K.* are non-speedy variants. In French, only the variant *O.K.* is conformant[8]. Therefore, the variants *okay*, *Okay*, *ok*, *OK*, *Ok*, *o.k.*, *O. K.* and *o. k.* were classified as being non-conformant.
(3) We integrated the French CMC sample in our cross-modal study on positional differences between spoken and written CMC, described in section 3.2., to investigate the distribution patterns in the CMC data of both languages.

4.3 Results and Discussion

(1) The results in Figure 3 support our assumption that OKAY is less frequently used as a syntactic unit in French than in German[9].

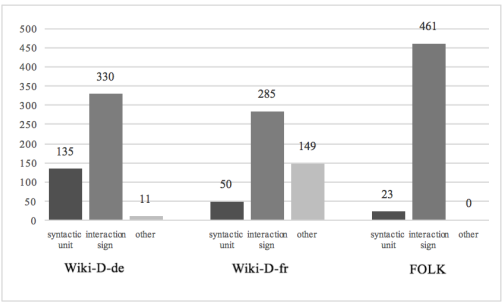

Figure 3: Functional distribution of OKAY.

In the German CMC data, 135 (28.4 %) occurrences of OKAY were tagged as syntactic units and 330 (69.3 %) occurrences as interaction signs. In French, 285 (58.9 %) occurrences were classified as interaction signs whereas 50 (10.3 %) were tagged as syntactic units. The French data included a considerably high amount of OKAY occurrences that could not clearly be classified as either interactive or syntactic (149 occurrences; 30.8 %)[10].
The aforementioned results had been achieved by manually checking and tagging the samples. Using a tagger that automatically assigns part-of-speech (POS) tags to distinguish between interactive or grammatical usages of OKAY did not achieve satisfactory results. The applied taggers either tagged all occurrences as being interactive, e.g. in FOLK, or as being grammatical, e.g. in Wiki-D-de. Studying OKAY exemplifies that there is still a need for improvement in the field of POS-tagging (cf. Lüngen et al., 2016 for details).
(2) The results of our cross-lingual study on the frequency of spelling variants are presented in Table 2 of section 2.2. The most frequent variants in the corpora are non-conformant, but support speed-writing. In both languages the non-speedy variants including a space (*o. k.* and *O. K.*) are rarely used. The variants *okay* and *Okay* are less frequent in French than in German, where these forms are norm-conformant.
(3) In terms of the positional distribution, shown in Figure 1, there is a clear distinction between speech and written CMC corpora. The distributional patterns in the French and the German CMC data do not differ to a vast extent and therefore seem to be language independent.

5. Conclusion

We investigated the usage of OKAY across genre types (German Wikipedia articles vs. talk pages), modes (German spoken vs. written interaction), and across languages (German vs. French CMC). The cross-genre study illustrated that OKAY is quite more frequently used in the CMC genre and that speedy writing variants are preferred over rule-conformant non-speedy ones. The cross-lingual study revealed that the grammatically integrated functions of OKAY occur more frequently in the German than in the French data. This may be an effect of the French language policy that recommends to avoid English loan elements. By comparing the frequency of spelling variants we found that the "speedy" variants are highly preferred in French and in German, although these variants are not rule-conformant. The cross-modal study showed that the function of a responsive, described as being the main function in the GDS grammar, is rarely used in both written and spoken corpora. It is thus worthwhile to investigate the functions of OKAY on the basis of corpus data. The results of the comparison of positional categories in Figure 2 revealed that the distribution patterns in the French and the German CMC corpora are quite similar, whereas the patterns in the CMC corpora differ considerably from the distribution in the spoken language corpus FOLK. Further work will study the usage of interaction signs in spoken and written CMC interaction on the basis of a more fine-grained annotation of functional categories.

[8] Cf. Le Petit Robert, 2017 p. 1736.
[9] The samples contain absolute frequencies of true positives.
[10] Examples are posts containing elliptical constructions like "Donc, OK pour moi" or "OK pour la date de la mort".

6. References

Bangerter, A.; Clark, H.H.; Katz, A.R. (2003). Navigating Joint Projects in Telephone Conversations. In *Discourse Processes* 37, pp. 1-23.

Beach, W. (1993). Transitional regularities for 'casual' "Okay" usages. In *Journal of Pragmatics* 19, pp. 325-352.

Beißwenger, M. (2008). Situated Chat Analysis as a Window to the User's Perspective: Aspects of Temporal and Sequential Organization. In J. Androutsopoulos, M. Beißwenger (Eds.), *Data and Methods in Computer-Mediated Discourse Analysis* (= Language@Internet 5).

Beißwenger, M.; Ermakova, M.; Geyken, A.; Lemnitzer, L.; Storrer, A. (2012). A TEI Schema for the Representation of Computer-mediated Communication. In *Journal of the Text Encoding Initiative (jTEI)*. Issue 3/2012 (DOI: 10.4000/jtei.476).

Delahaie, J. (2009). Oui, voilà ou d'accord? Enseigner les marqueurs d'accord en classe de FL. In *Synergies Pays Scandinaves* 4, pp. 17-34.

Duden-Rechtschreibung (2013). *Duden – Die Grammatik*. 26., völlig neu erarbeitete und erweiterte Auflage. Berlin: Bibliographisches Institut GmbH. (= Band 1 – Der Duden in 12 Bänden).

Condon, S.L.; Čech, C.G. (2007). OK, next one: Discourse markers of common ground. In A. Fetzer, K. Fischer (Eds.), *Lexical Markers of Common Grounds*. London: Elsevier, pp. 18-45.

GDS (1997). *Grammatik der deutschen Sprache*. Zifonun, G.; Hoffmann, L.; Strecker, B.; et al. (Eds.). 3 Bände. Berlin/New York: de Gruyter.

Herzberg, L. (2016). *Korpuslinguistische Analyse interaktiver Einheiten: das Beispiel okay*. Master thesis. University of Mannheim.

Kaiser, J. (2011). *okay in ärztlichen Gesprächen – eine linguistische Gesprächsanalyse*. State examination thesis. Ruprecht-Karls-University Heidelberg.

Le Petit Robert (2017). *Dictionnaire Alphabétique Et Analogique De La Langue Française*. Paris: Dictionnaires Le Robert.

Levin, H.; Gray, D. (1983). The Lecture's OK. In *American Speech* 58, pp. 195-200.

Lüngen, H.; Beißwenger, M.; Herold, A.; Storrer, A. (2016). Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In S. Dipper, F. Neubarth, H. Zinsmeister (Eds.), *Proceedings of the 13th Conference on Natural Language Processing* (KONVENS 2016), pp. 156-164.

Margaretha, E.; Lüngen, H. (2014). Building linguistic corpora from Wikipedia articles and discussions. In *Journal of Language Technologie and Computational Linguistics JLCL* 29(2), pp. 59-83.

Schegloff, E.A.; Sacks, H. (1973). Opening up Closings. In *Semiotica* 8, pp. 289-327.

Schegloff, E.A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen (Ed.), *Analyzing discourse: Text and talk*, Washington, DC: Georgetown University Press, pp. 71-93.

Storrer, A. (2017). Grammatische Variation in Gespräch, Text und internetbasierter Kommunikation. In M. Konopka, A. Wöllstein (Eds.), *Grammatische Variation. Empirische Zugänge und theoretische Modellierung*, Berlin/New York: de Gruyter, pp. 105-125.

Corpus tools and resources:

COSMAS I/II: Corpus Search, Management and Analysis System. Institute for the German language Mannheim, http://www.ids-mannheim.de/cosmas2/.

DeReKo: Das Deutsche Referenzkorpus. Institute for the German language Mannheim, http://www.ids-mannheim.de/kl/projekte/korpora/.

DGD: Datenbank gesprochenes Deutsch. Institute for the German language Mannheim, http://agd.ids-mannheim.de/folk.shtml.

FOLK: Forschungs- und Lehrkorpus für gesprochenes Deutsch. Institute for the German language Mannheim, http://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome.

RAPIDMINER-KOBRA: RapidMiner Software, www.rapidminer.com and KobRA Plugin, http://www.kobra.tu-dortmund.de/mediawiki/index.php?title=Software.

Wiki-A-de: Corpus with all articles of the German Wikipedia (Version 17.11.2015). Institute for the German language Mannheim, http://corpora.ids-mannheim.de/pub/wikipedia-deutsch/2015/.

Wiki-D-de: Corpus with all article talk pages of the German Wikipedia (Version 17.11.2015). Institute for the German language Mannheim, http://corpora.ids-mannheim.de/pub/wikipedia-deutsch/2015/.

Wiki-D-fr: Corpus with all article talk pages of the French Wikipedia (Version 17.11.2015). Institute for the German language Mannheim, http://corpora.ids-mannheim.de/pub/wikipedia-fremdspr/2015/.