On the Role of Historical Newspapers in Disseminating Foreign Words in German

Oliver Pfefferkorn, Peter Fankhauser

IDS-Manneim
Germany
pfefferkorn@ids-mannheim.de, fankhauser@ids-mannheim.de

Abstract

Newspapers became extremely popular in Germany during the 18^{th} and 19^{th} century, and thus increasingly influential for modern German. However, due to the lack of digitized historical newspaper corpora for German, this influence could not be analyzed systematically. In this paper, we introduce the Mannheim Corpus of Digital Newspapers and Magazines, which in its current release comprises 21 newspapers and magazines from the 18^{th} and 19^{th} century. With over 4.1 Mio tokens in about 650 volumes it currently constitutes the largest historical corpus dedicated to newspapers in German. We briefly discuss the prospect of the corpus for analyzing the evolution of news as a genre in its own right and the influence of contextual parameters such as region and register on the language of news. We then focus on one historically influential aspect of newspapers – their role in disseminating foreign words in German. Our preliminary quantitative results indeed indicate that newspapers use foreign words significantly more frequently than other genres, in particular belle lettres.

Keywords: Historical Corpora, Newspapers, Language Variation

1. Introduction

Newspapers became extremely popular in Germany during the 18^{th} and 19^{th} century. In the 18^{th} century several types of newspapers, among them political newspapers, advertising supplements or intelligencers, and weekly magazines emerged, coinciding with the rise of a civil public. In contrast to other types of public and private writing, newspapers are characterized by actuality, periodicity, topical diversity, and public access (Wilke, 1999, p. 388). They reached a wider audience than any other text type and spread across all social classes.

Wilke (1999, p. 397) estimates that German newspapers reached about 3 Mio readers in the late 18^{th} century. Thus, the language of newspapers has been very influential for the development of German in the 18^{th} and 19^{th} century. Up to now this influence has not been analyzed systematically; rather analysis of the more recent history of German has focussed on few selected varieties, in particular (high) literature.

A historical reason for this marginalization of newspapers as an influential factor for (early) modern German is the rather widespread contempt for the medium by the educated middle class (Theobald, 2012). A more practical reason though is the unavailability of newspaper corpora. Newspapers typically have not been reprinted, and the original facsimiles are difficult to access. Therefore, research on language history is typically focussed on the analysis of individual newspapers and rather specific linguistic questions.

The remainder of this paper is organized as follows: In Section 2. we describe the Mannheim Corpus of Historical Newspapers and Magazines and in Section 3. we outline some of its potential uses for analyzing the influence of newspapers on modern German. In Section 4. we analyze the role of newspapers in popularizing foreign words, and in Section 5., we conclude and outline future work.

2. The Mannheim Corpus of Historical Newspapers and Magazines

The Mannheim Corpus of Historical Newspapers and Magazines (MKHZ) consists of 21 German newspapers and magazines from the 18^{th} and 19^{th} century. The currently publically available version[1] comprises about 650 individual volumes with over 4.1 Mio tokens on 4678 pages overall[2].

In addition to the original page scans, it is available in TUSTEP format (TUSTEP, 2013), acquired in a double keying procedure, and in TEI P5, which was generated semi-automatically from the TUSTEP version (Fankhauser et al., 2013b). On this basis, a human readable version in HTML aligned with the page scans, CMDI metadata (Broeder et al., 2011) for its long term archival in the IDS Repository (Fankhauser et al., 2013a), and an *IDS–XCES* version for import into the Corpus Search and Analysis System *COSMAS II* (Bodmer, 2005) were generated.

For each volume, basic metadata including title, publication date and place, and a broad classification into newspapers vs. magazines are available. The logical structure is very simple, individual volumes consist of paragraphs and tables, a division into individual articles is currently underway.

3. Digitized Newspapers for Linguistic Research

Digitized newspaper corpora enable a quantitative perspective on the role of news for the historical development of German. Here is a nonexhaustive list of potentially interesting questions to investigate:

[1] *MKHZ*: http://hdl.handle.net/10932/00-01B8-AE41-41A4-DC01-5

[2] For comparison, the historical newspaper corpus compiled for the GerManC corpus (Bennett et al., 2010) comprises about 300.000 tokens.

News as a Genre: How did the news genre differentiate itself from other genres, such as science or fiction?

Types of News: How did various types of news and newspapers evolve and differentiate themselves (e.g. news, announcement, background feature, reportage (Püschel, 1999))? And how do these types manifest themselves linguistically?

Homogenization: Whether and how were regional and dialectal characteristics reduced over time?

Advertisements: How did various types of advertisements evolve (e.g. product ads, job ads, real estate ads, private ads, personal ads, purchase ads, etc.)?

Spoken Language: Newspapers contain text types pertaining to some extent to spoken language, such as speeches, protocols, calls, or letters to the editor, which may help in analyzing historical spoken language.

4. Foreign Words in Newspapers

In this section we analyze one aspect of newspapers as a genre in more detail. Newspapers and magazines have never been shy of using foreign words. This can already be observed for the early German weekly magazines in the 17^{th} century (Gloning, 1996, pp. 164 – 179). Authors of dictionaries of foreign words in the 18^{th} and 19^{th} almost formulaically refer to "assisting their readers in comprehending newspapers" in their subtitles. This gives rise to the hypothesis that newspapers took over and popularized many foreign words from specialized subject domains. As a consequence many of these foreign words have been taken over into standard German.

Until now though, the historical role of newspapers in disseminating foreign words has not been analyzed systematically due to the lack of appropriate historical newspaper corpora. It is unclear which foreign words were used most frequently in newspapers, which foreign words in contemporary dictionaries of the 18^{th} and 19^{th} century were used at all, and what their percentages in comparison to other genres or registers were.

For a preliminary investigation of these questions, we have derived a list of foreign words from a German dictionary of foreign words, originally started in 1913 and continually revised since then[3]. The list comprises about 3700 *main* lemmata for the letters "A" to "Q". It is by no means exhaustive; it does not cover letters "R" to "Z", it only takes into account the main lemmata, and it does not cover foreign words that have become obsolete again. Nevertheless, it can be regarded as a representative sample for foreign words that have been considered part of standard German by lexicographers [4]

For comparing the historical role of newspapers in picking up foreign words as opposed to other registers, we use in addition the German Text Archive *DTA*[5] (Geyken et al., 2011), compiled as a representative cross section of texts from 1600 – 1920. *DTA* currently comprises about 90 Mio tokens, and is drawn from the following three broad genres: factual writing (*Gebrauchsliteratur*), belle lettres (fiction, drama, poetry etc.), and learned (scientific writing), but has only a relatively small coverage of historical newspapers.

From both corpora a list of all types together with their frequencies is extracted. To match inflected forms with the lemmata from the list of foreign words, they are lemmatized with a word based lemmatizer (Belica, 1994) for German, and normalized using a small, non-exhaustive set of heuristic rules. The resulting matches are not 100% accurate; ambiguous words such as *modern* (as a (foreign) adjective: *modern*, as a (native) verb: *to molder/rot*) hurt precision, incomplete normalization and lemmatization, and not taking into account derived lemmata hurt recall. However, these inaccuracies have largely the same effect on the various (sub)corpora; fairly low recall is the dominant factor, and thus the reported percentages of foreign words in German constitute lower bound estimates.

Table 1 lists the overall number of tokens (*#t*), the number of tokens which match foreign words (*#f*), and the according percentage of foreign words (*%f*) for *MKHZ* and *DTA* over time[6] The relatively low percentage of foreign words in the 17^{th} century (*DTA* only) is mainly due to two reasons: *DFWB* does not include foreign words that have become obsolete again, and normalization and lemmatization are less accurate for this period. However, in the subsequent periods, which are covered by both corpora, the percentage of foreign words in *DTA* is consistently smaller than in *MKHZ*[7]. This indicates that indeed newspapers have historically contributed more strongly to disseminating foreign words into general language than other genres.

Table 2 analyzes the percentage of foreign words by genre. *MKHZ* is broadly divided into newspapers vs. (weekly) magazines, which often contain text pertaining more to belle lettres, such as serial novels. Indeed newspapers proper have a significantly higher percentage of foreign words than magazines. This is mirrored in *DTA* where foreign words in factual writing are also significantly more frequent than in belle lettres. As is to be expected scientific writing (*learned*) has the highest frequency of foreign words in *DTA*. Note that newspapers in *MKHZ* also have a higher percentage than scientific writing in *DTA*, but this may again well be due to the longer time span covered by *DTA*.

For a more detailed perspective, we analyze the dissemination of foreign words during the 18^{th} and 19^{th} century along three examples. In addition to *MKHZ* and *DTA*, we also use the historical corpus *DGB01*[8], which comprises

[3]Deutsches Fremdwörterbuch (DFWB). http://www1.ids-mannheim.de/lexik/fremdwort/

[4]Main lemmata in *DFWB* comprise loan words as well as some words derived by combining loan stems, prefixes, and suffixes with other words (derived lemmata).

[5]Deutsches Text Archiv. http://www.deutschestextarchiv.de. Version: Nov. 6, 2013, downloaded Feb. 11, 2014.

[6]The overall number of tokens is smaller for both corpora, because only alphabetic words occurring within paragraphs – no titles, tables, etc. – have been taken into account.

[7]All reported differences are significant according to a χ^2 test with a p-value well below 0.1%.

[8]Deutsche Bibliothek/Deutsche Literatur von Lessing bis Kafka (Digital Library/German Literature from Lessing to Kafka)

		MKHZ	DTA
1600 – 1700	#t	–	8176835
	#f	–	86664
	%f	–	1.06
1700 – 1800	#t	202654	14743225
	#f	5071	211507
	%f	2.50	1.43
1800 – 1850	#t	1548983	18565922
	#f	31168	348225
	%f	2.01	1.88
1850 – 1920	#t	2161855	28872818
	#f	51965	652814
	%f	2.61	2.26
Sum	#t	3913492	70358800
	#f	88222	1299210
	%f	2.25	1.85

Table 1: Foreign Words along Time

		MKHZ	DTA
newspapers /	#t	2293010	7724113
factual writing	#f	59924	140250
	%f	2.61	1.82
magazines /	#t	1620482	16928838
belle lettres	#f	28298	216876
	%f	1.75	1.28
learned	#t	–	45705849
	#f	–	942084
	%f	–	2.06

Table 2: Foreign Words by Genre

about 30 Mio tokens.

The noun *Agentur* (*agency*) was defined in dictionaries of the late 18^{th} century as a derivation from *agent* meaning *office, capacity, assignment by an agent, mediator, representative*[9]. The revised *DFWB* (Strauß et al., 1995, p. 192) reports its first record in 1847 with the meaning *branch office, office of an agent*[10]. Since the beginning of the 19^{th} century *Agentur* occurs regularly, but only since about 1870 it also occurs frequently. *DGB01* has only 3 occurrences, whereas *MKHZ* lists 149 occurrences between 1846 and 1877. In addition, *MKHZ* lists 30 occurrences of derived composites, such as *Generalagentur, Generalzeitungsagentur, Gesellschaftsagentur, Hauptagentur, Nachrichtenagentur, Patentagentur, Spezialagentur*. *DTA*, has only 32 occurrences of *Agentur* between 1855 and 1900. This strongly suggests that mainly newspapers have contributed to establishing *Agentur* in standard German.

The first edition of the *DFWB* reports 1779 as the first record of the adjective *provisorisch* (*provisional, temporary*) in German (Schulz and Basler, 1942, p. 716). Until the mid 19^{th} century, it occurs almost exclusively in scientific contexts, in particular legal and philosophical writing, and only rarely in literary publications, as also evidenced by *DTA*. Only since the mid 19^{th} century it occurs more frequently outside of scientific writing, and indeed primarily in newspapers and magazines (367 occurrences in MKHZ between 1847 and 1905 with a peak between 1848 and 1850). It is typically used in political contexts, such as in *der provisorische Zustand (the provisional state of affairs)*, *die provisorische Regierung/Zentralgewalt (the provisional government/central power)*, or *provisorische Anordnungen/Gesetze (provisional orders/laws)*. In comparison, the literature corpus *DGB01* only records 53 occurrences in the period of 1793 and 1923. In this case, newspapers adopt the scientific terminology, adapt it to more domains, and thereby disseminate it.

As a last example, the adjective *reaktionär* (*reactionary*) was borrowed from French in the eighteen-thirties. Until the middle of the 18^{th} century it was used sparsely in political writing (10 occurrences in *DTA* from 1833 to 1849). Only since 1855 it started occurring more frequently in more domains. This increase is paralleled in *MKHZ*, where its frequency reaches a peak between 1848 and 1852 with 51 occurrences, with 73 occurrences overall. This again suggests that the language of newspapers may have served as a mediator, disseminating a specialized foreign word into standard German. In comparison, *DBG01* contains only 26 occurrences between 1848 and 1905.

5. Conclusions and Future Work

We have presented the Mannheim Corpus of Historical Newspapers and Magazines, and on this basis investigated the role of newspapers for popularizing foreign words in German. Currently we are working on transforming another 2500 pages from TUSTEP to TEI P5 to improve its coverage. Moreover, we cooperate with the Berlin Brandenburg Akademie der Wissenschaften to integrate the corpus into *DTA*, which will enable us to use the more advanced techniques to normalization employed by the *DTA* for our empirical analysis.

Lexical change is not confined to the introduction of foreign words. In the course of the lexical and semantic change in German during the early 19^{th} century many words from middle and early modern German – in particular in poetic literature – became obsolete or changed their meaning (Beutin, 1972), leading to the final norm of modern German also in the lexical domain. So far the reasons for this change have not been described systematically, e.g., which were the driving social forces, or in which genres and registers did this change occur first. Like with respect to foreign words, newspapers as a fairly new medium could have played an important role in this change. To help tracking possible semantic innovation in newspapers, we plan to systematically analyze and compare the local contexts of words in various registers and over time.

Finally, we also plan to compare the historical newspaper corpus with contemporary newspapers to get a better understanding of the evolution of news as a genre, looking at effects of conventionalization and diversification w.r.t. other genres.

[9] Amt, Funktion, Auftrag eines Bevollmächtigten, Vermittlers, Vertreters

[10] Geschäftsstelle, Büro eines Agenten

Acknowledgements

The curation of MKHZ has been partially supported by DFG (Deutsche Forschungsgemeinschaft; area: Wissenschaftliche Literaturversorgungs- und Informationssysteme, project: Zentrum für germanistische Forschungsprimärdaten).

6. References

Cyril Belica. 1994. WP2 – Lemmatizer, Final Report. Technical report, Institut für Deutsche Sprache, July.

Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt. 2010. Annotating a historical corpus of german: A case study. In *Proceedings of the LREC 2010 workshop on Language Resources and Language Technology Standards*, pages 64–68, Valletta, Malta, May.

Wolfgang Beutin. 1972. *Das Weiterleben alter Wortbedeutungen in der neueren deutschen Literatur bis gegen 1800.* Lüdke, Hamburg.

Franck Bodmer. 2005. COSMAS II. Recherchieren in den Korpora des IDS. *Sprachreport*, 3:2–5.

Daan Broeder, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck, and Andreas Witt. 2011. A pragmatic approach to XML interoperability - the Component Metadata Infrastructure (CMDI). In *Proceedings of Balisage : The Markup Conference 2011*, volume 7. Balisage Series of Markup Technologies.

Peter Fankhauser, Norman Fiedler, and Andreas Witt. 2013a. Forschungsdatenmanagement in den Geisteswissenschaften am Beispiel der germanistischen Linguistik. *Zeitschrift ZfBB, Zeitschrift für Bibliothekswesen und Bibliographie*, 60, December.

Peter Fankhauser, Oliver Pfefferkorn, and Andreas Witt. 2013b. From TUSTEP to TEI in Baby Steps. In Fabio Cotti and Arianna Ciula, editors, *Abstracts of the TEI Conference and Members Meeting*, pages 34–38, Rome, October. DIGILAB Sapienza University & TEI Consortium.

Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas, and Frank Wiegand. 2011. Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In Silke Schomburg, Claus Leggewie, Henning Lobin, and Cornelius Puschmann, editors, *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland, 2010, 20./21.September*, pages 157–161. 2. ergänzte Fassung, hbz.

Thomas Gloning. 1996. Bestandsaufnahme zum Untersuchungsbereich "Wortschatz". In Gerd Fritz and Erich Straßner, editors, *Die Sprache der ersten deutschen Wochenzeitungen im 17. Jahrhundert*, pages 164 – 195. deGruyter, Tübingen.

Ulrich Püschel. 1999. Präsentationsformen, Texttypen und kommunikative Leistungen der Sprache in Zeitungen und Zeitschriften. In Joachim-Felix Leonhard, Hans-Werner Ludwig, Dietrich Schwarze, and Erich Straßner, editors, *Medienwissenschaft. Ein Handbuch zur Entwicklung der Medien und Kommunikationsformen*, volume 1, pages 865 – 868. deGruyter, Berlin, New York.

Hans Schulz and Otto Basler. 1942. *Deutsches Fremdwörterbuch. B2: L–P.* Walter de Gruyter, Berlin.

Gerhard Strauß, Elke Donalies, Heidrun Kämper-Jensen, Isolde Nortmeyer, Joachim Schildt, Rosemarie Schnerrer, and Oda Vietze. 1995. *Deutsches Fremdwörterbuch. Bd. 1, a-Präfix Antike. Völlig neubearbeitet im Institut für Deutsche Sprache*, volume XVII. de Gruyter, Berlin/New York, 2 edition.

Tina Theobald. 2012. "Dieses unselige Zeitungsdeutsch" Reflexion über die Presse und ihren sprachlichen Einfluss im 19. Jahrhundert. *Sprachreport*, (3):12 – 21.

TUSTEP. 2013. Handbuch und Referenz (electronic version). Technical report, Universität Tübingen; Zentrum für Datenverarbeitung.

Jürgen Wilke. 1999. Die Zeitung. In Ernst Fischer, Wilhelm Haefs, and York-Gothart Mix, editors, *Von Almanach bis Zeitung. Ein Handbuch der Medien in Deutschland 1700 - 1800*, pages 388 – 402. C.H. Beck, München.