Poster presented at the 'Doktorandentag 2011' of the department of German Language and Linguistics, Humboldt Universität zu Berlin, 12.10.2011

Tracing Sociocultural Change through Multi-Word Sequences

Andreas Bürki
University of Basel and Humboldt Universität zu Berlin

contact: andreas.buerki@unibas.ch
supervisor Basel: Annelies Häcki Buhofer
supervisor Berlin: Anke Lüdeling

❶ Background: Multi-Word Sequences

Multi-word sequences (MWSs) are *'semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments'* (Sinclair 1991:110). They are thought to:

- represent a quantitatively important part of language use (Altenberg 1998, Erman and Warren 2000)
- ease processing load in language production and reception, thus enabling fluency (Wray 2002 and others)
- activate a pragmatic-situational as well as cultural background and thus facilitate understanding of what is said or written (Feilke 1993)
- be central to the linguistic system (Sinclair 1991) but neither strictly lexis nor quite syntax

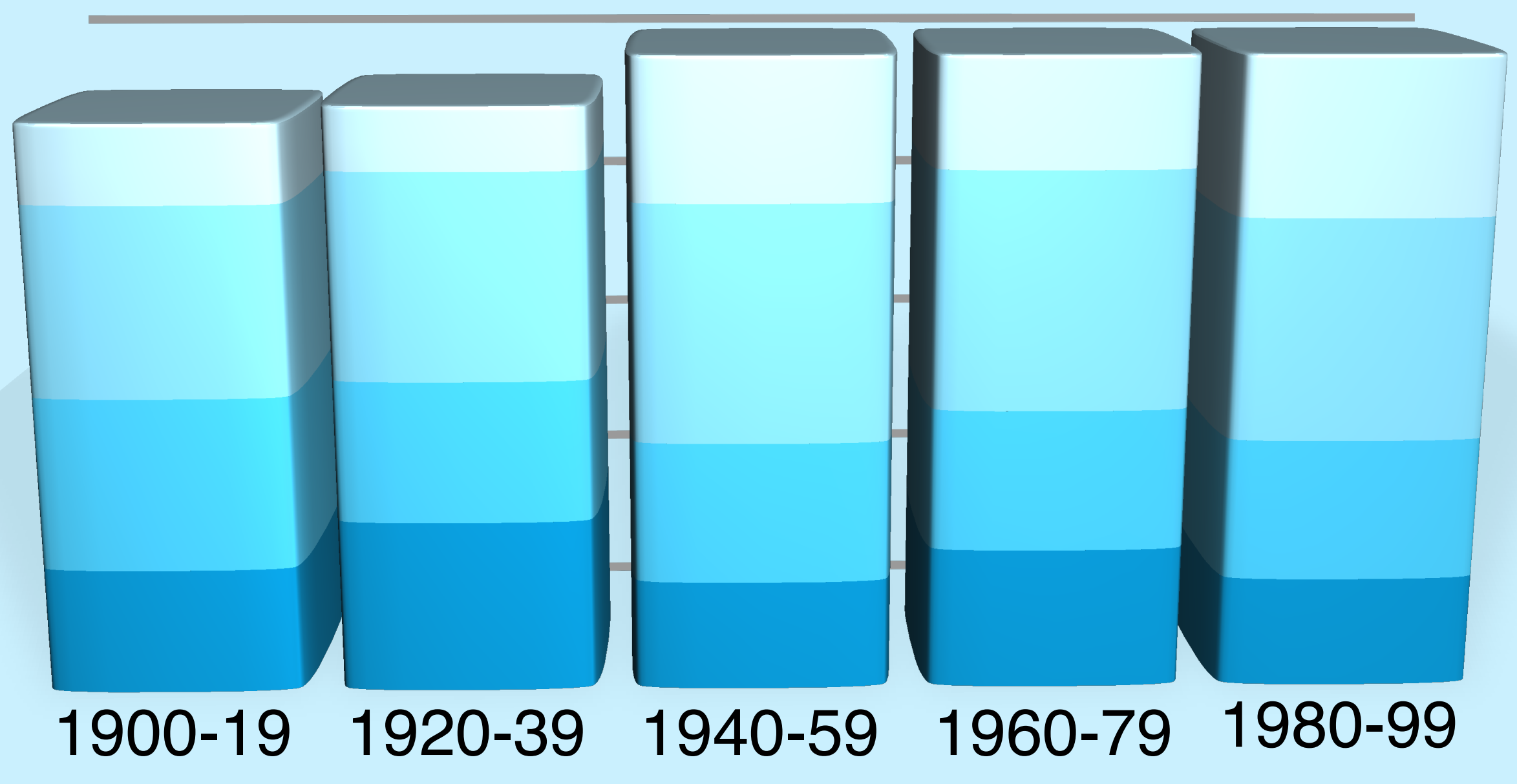
MWSs represent the way things are customarily put in a speech community. Though this would suggest links to the culture of the community, this has not on the whole received much scholarly attention, except in the case of particular types of MWSs like proverbs (as in Teliya et al.1998).

❷ Principal Question

To what extent are changes in MWSs related to cultural change?

❸ Data and Method

The data for this project are taken from the Swiss Text Corpus (http://chtk.unibas.ch), a 20-million word reference corpus of written standard German as used in the German-speaking part of Switzerland. It is balanced across four genre groups (journalistic prose, texts such as advertising that are written primarily for a targeted audience, subject texts and literature. The corpus covers language from 1900 to 2000 and is PoS-tagged and lemmatised.



For analysis, the corpus is split into 5 periods of 20 years each, resulting in 5 subcorpora of around 4M words.

Using a version of the NSP (Banerjee and Pedersen 2003; Wilmsmann 2007) and custom-made UNIX shell scripts, N-gram lists of length 2-7 are extracted and consolidated into one single list per time period. To extract MWSs from N-grams, a cutoff at 4 occurrences per million words and occurrence in at least 2 documents is enforced. A further automatic filter is then applied, resulting in an extraction accuracy of around 70% on lists (Bürki, forthcoming).

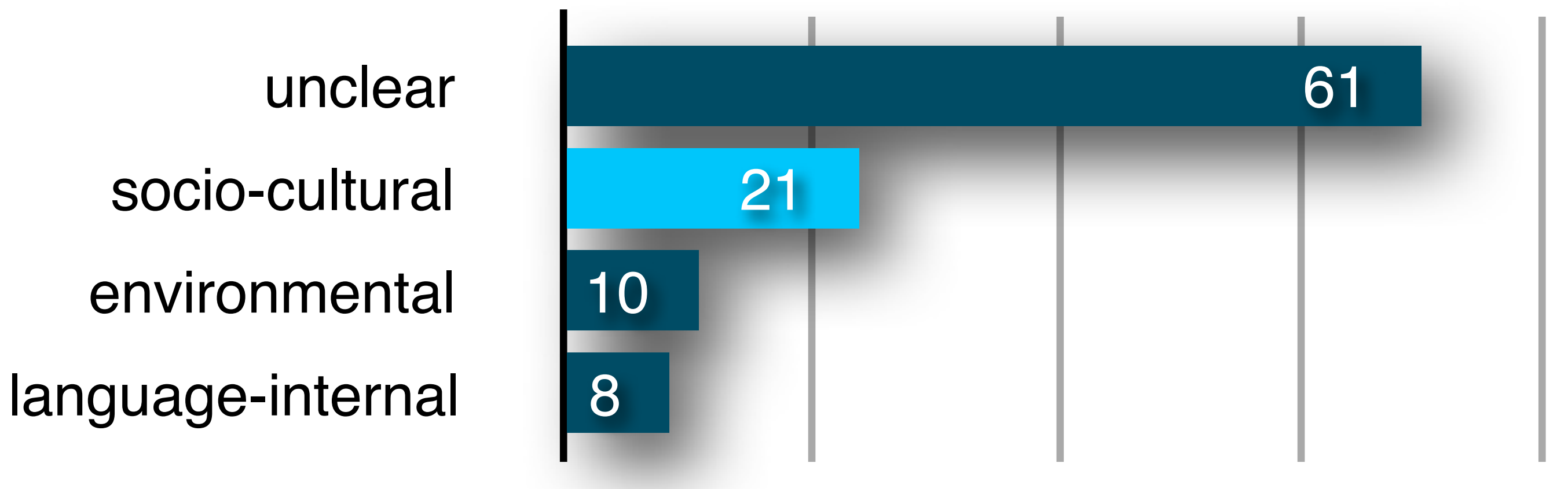
Lists for each of the 5 periods are then compared and MWSs which display frequency differences between periods are assembled into a list of changes. An example entry is shown on the right.

MWS	1900-19	1920-39	1940-59	1960-79	1980-99
aus diesem Grund	-	-	-	30	44
aus diesem Grunde *[for this reason]*	49	68	89	54	29

To assess the extent to which observed changes in MWSs relate to cultural change, the list of changes is analysed, changes categorised, motivating factors researched and regularities and wider implications investigated.

❹ Some early results

In a random sample of 100 observed changes, each change was assigned one of 4 categories of motivation: 1) unclear (assigned when motivation remained unclear after 10 minutes of research on a particular change) 2) sociocultural 3) environmental (natural, technological and built) 4) language-internal. The assignment of changes to these categories was done making use of concordance lines and general knowledge. The proportion of changes in MWSs linked to sociocultural change was 21% in the sample:



above: motivations for changes in a random sample of all changes (N=100).

Examples of changes in MWSs motivated by cultural change:

MWS	1900-19	1920-39	1940-59	1960-79	1980-99
ⓐ **die Forderung nach** *[the demand for (a right)]*	**-**	**-**	**25**	**49**	**41**
ⓑ**blauen Augen** *[blue eyes]*	**33**	**25**	**19**	**-**	**-**
ⓒ **der Glaube an** *[the belief/faith in]*	**28**	**24**	**17**	**-**	**-**
ⓓ **der liebe Gott** *[the good Lord]*	**46**	**41**	**22**	**21**	**-**
ⓔ **in unserem Land[e]** *[in our country]*	**23**	**23**	**99**	**80**	**90**

ⓐ The slot at the end of this MWS is very frequently filled with words denoting concepts like *more rights* and *better conditions of various sorts*. Peaking in the 60s and 70s, this MWS appears to have come into existence as a MWS through the social movements today referred to by the iconic year of 1968.

ⓑ Concordance lines for this MWS show that in a large number of cases, mention is also made, in the vicinity, of blond hair. This type of description appears to have become distinctly unfashionable even in German-speaking Switzerland with the end of the Second World War and the demise of fascist racial ideology.

ⓒ+ⓓ have to do with religion - they decline and disappear as MWSs over the 20th century. This may well reflect the loss of the importance of religion in the public sphere over the period. Other MWSs that show a closely similar development include *im Himmel [in heaven]* and *in der Seele [in the soul]*.

ⓔ The frequency of this MWS shoots up at the height of World War 2, a pattern mirrored in the data by the adjective *schweizerisch [Swiss] among others*. The popularity of this MWS appears rather directly related to the programme of state-sponsered nationalism referred to as *geistige Landesverteidigung* instigated at that same time by the Swiss government in an attempt to aid national defence.

❺ Conclusions and Outlook

Early results demonstrate a link between cultural change and frequency changes in MWSs in the data. The extent to which the two are related appears notable, despite the difficulties in establishing clear motivations for many of the observed MWS-changes. Examples show that the relationship between cultural change and changes in MWS is not limited to certain topical areas, but encompasses diverse domains. While more detailed analyses of observed changes will be needed to confirm and add detail to the initial results presented here, it is clear that sociocultural change is an important motivating factor for linguistic change beyond the areas of lexical change and the spread of (sound-)change. This represents a departure from current thought in historical linguistics.

References: **Altenberg**, B. (1998) "On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations." In *Phraseology: Theory, Analysis and Applications*. Edited by A. Cowie. Oxford: Clarendon Press. **Banerjee**, S. and T. **Pedersen** (2003) "The Design, Implementation and Use of the Ngram Statistics Package." In *Proceedings of the 4Th International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City.* **Bürki**, A. (fourthcoming) "Korpusgeleitete Extraktion von Mehrwortsequenzen aus (diachronen) Korpora: Vorgehenswege für deutschsprachige Daten" in *Aspekte der historischen Phraseologie und Phraseographie* edited by Filatkina et al. **Erman**, B, and B **Warren** (2000) "The Idiom Principle and the Open Choice Principle." *Text* 20, no. 1: 29-62. **Feilke**, H. (1993) "Sprachlicher Common Sense Und Kommunikation. Über Den gesunden Menschenverstand, Die Prägung Der Kompetenz Und Die Idiomatische Ordnung Des Verstehens." *Der Deutschunterricht* 45, no. 1993: 6-21. **Sinclair**, J (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press. **Teliya**, V., N. Bragina, E. Oparina, and I. Sandomirskaya (1998) "Phraseology As a Language of Culture." In *Phraseology: Theory, Analysis, and Applications*. Edited by A P Cowie. Oxford: Clarendon Press. **Wray**, A.(2002) *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press. Wilmsmann, B. (2007) "Re-Write of Text-NSP". http://topicalizer.com/files/TextNSP/Re-write_of_Text-NSP.pdf .