Steffen Eger

Investigating lexical competition — An Empirical Case Study of the German Spelling Reform of 1996/2004/2006

1 Introduction

The German spelling reform of 1996/2004/2006 triggered the introduction of new orthographic variants in the German spelling system. These were the products of different kinds of modifications enacted by the reform. They could be a result of a 'mutation'-like change of some of the characters of a word (as, for example, the change from *Biographie* to *Biografie*), due to a writing as two words of a word form formerly written as one word (as in *kennen lernen* vs. *kennenlernen*), due to the introduction of a hyphenation (as in *17-jährig* vs. *17jährig*) or due to a change in the lower or upper case writing of words (as in *im Allgemeinen* vs. *im allgemeinen*). The goal of the current study is to present a transferable methodological framework in which the developments of the German spelling reform can be studied — more precisely, the reactions of the language users, as representable by language corpora, to the specifications purported by the reform. Particular interest lies in the *distribution of competing forms*; the spelling reform in general caused the simultaneous co-existence of two or, occassionally, more (semantically equivalent) forms, and the current survey tries to sketch the relative status of these competitors over time.

The methods of analysis we thereby choose are general enough to be not only applicable to the particular situation of the German spelling reform, but to every state of affairs where two linguistic features are (partially) synonymous and are hence strict alternatives ("competitors") of which the language user may choose. This encompasses for example the competition of a 'native' and a 'foreign' form in a particular natural language — for example, in German, many modern English words are rivalling with traditional forms such as *user* vs. *Benutzer*, *Band* vs. *Gruppe*, etc. — or the competition of other alternatives of varying origins such as in German indicative imperfect *gewänne* vs. *gewönne*, *stünde* vs. *stände*, etc., in English past participle *shown* vs. *showed*, simple past *dreamed* vs. *dreamt*, etc. or as in British versus American English *labour* vs. *labor*, *bath* vs. *bathe*.

The structure of the current work is as follows. In Section 2 we give a short introduction to the German spelling reform and the changes in the German orthographic system it entailed. Section 3 presents an overview over the data we use, which is based on DeReKo, the German reference corpus at the Institute for the German Language (IDS). Before illustrating the results of our analysis in Section 5, we detail various aspects of our methodological approach in Section 4; these comprise besides a time series representation of our data principal component analyses and clustering techniques for

evaluation and generalization. After a short discussion in Section 6, which focuses on the reformed language features accepted and not accepted by the language community, we conclude in Section 7.

2 The German Spelling Reform of 1996/2004/2006

The major goal of the German Spelling Reform of 1996/2004/2006 was a simplification of the rules underlying the German spelling system in order to adapt it to modern standards (IAO (1992)). The reform was implemented in three stages; the main reform of 1996 was supplemented/revised by the 2004 and 2006 regulations, primarily in order to address the various forms of criticism brought forward against the original reformation.

The reform addressed roughly six major aspects of the German spelling system; in the following, we will shortly describe these. Our illustration will, however, be rather short and summarizing and we will not separately address individual reformations regulated by particular stages of the reform, but just give a generalizing overview. For a more detailed exposition we refer to the respective literature (e.g. Güthert (2006), Korrekturservice im Internet (2010), etc.).

(i) **Alignment of sounds and letters (ASL)**. Most prominently, this concerned the usage of *-ss-* and *-ß-*, but also the writing of particular foreign words. The following examples, where each instance represents a pre-reform/post-reform word pair,[1] illustrate some of the implemented reformations: *Fluß/Fluss*, *stillegen/stilllegen*, *Babies/Babys*, *Differential/Differenzial*, *numerieren/nummerieren*, *aufwendig/aufwändig*, *Biographie/Biografie*, *Joghurt/Jogurt*, *Spaghetti/Spagetti*. Some of these reformations were made mandatory (e.g. *Fluss* was to replace *Fluß*), while others were to be optional alternatives, at the disposal of the language user (e.g. *Spaghetti/Spagetti*).

(ii) **Writing as one or two words (W12)**. Here, the most radical modification of the 1996 reform was the consistent writing as two words of verb-verb combinations such as *sitzenbleiben/sitzen bleiben*, *kennenlernen/kennen lernen* etc., independent of metaphorical or concrete meaning (e.g. in prereform usage: *sitzen bleiben* "to remain seated" vs. *sitzenbleiben* "to stay down a year"). Also in many other cases like many combinations of particles and verbs such as *abwärtsfahren/abwärts fahren*, writing as two words was to replace writing as one word.

(iii) **Hyphenation**. Here, the reform generally prescribed the use of hyphens in situations of e.g. compounds involving numbers, abbrevations, etc. such as *17jährig/17-jährig*, *Email/E-Mail*, and so on. On the other hand, for angliscism compounds such as *Midlife-crisis/Midlifecrisis*, the spelling without a hyphen was to be allowed. In general, however, a more frequent use of hyphens was recommended, particularly for reasons of clarity, as in *Kaffeeextrakt/Kaffee-Extrakt*.

[1] Occasionally we will refer to such a pair by the post-reform variant solely.

(iv) **Lower and upper case (LUC)**. The idea of the reform here was to give formal rules for the use of lower and upper case. Hence, in particular, nominalizations involving articles such as *im folgenden/im Folgenden* were to be capitalized. Also, times of the day in combinations with *gestern, heute, morgen* such as *gestern abend/gestern Abend* were to be capitalized. The same was true for adjectival doublets like *leid tun/Leid tun, recht haben/Recht haben*, etc. On the other hand, in fixed combinations of adjectives and nouns with proper name character such as *Schwarzes Brett/schwarzes Brett, Erste Hilfe/erste Hilfe* the adjective was supposed to be lower case.

(v) **Punctuation**. This involved i.a. a simplification of comma placement rules, allowing the individual language user more freedom.

(vi) **End-of-line word separation**. Here, i.a., the rule of leaving *st* unseparated was abolished; e.g. analogously to the separation *Wes-pe* it was now allowed to separate *Weste* as *Wes-te*.

3 Data

Our data base is the IDS DeReKo (Kupietz and Keibel (2009), Institut für Deutsche Sprache (2010)) archive of written language. DeReKo represents the world-wide largest collection of electronically availabe corpora in the German language.[2] While it also comprises texts from science and fiction, its major component is newspaper corpora. In the current study, we focus exclusively on this last element of DeReKo because of the scarcity of the other resources. For the same reason, the time period we consider is restricted to the years 1985 to 2009; since the spelling reform took place in 1996 (respectively 2004 and 2006), this time frame should be suitable for making adequate statements with regard to the evolution of the reform. For the considered period, there are 31 different newspapers in DeReKo (including *Die Zeit*, *Mannheimer Morgen*, *taz*, *FAZ*, etc.), none of which is chronicled in every year. In fact, there are on average only about 9 newspapers for any given year in the epoch under analysis, with more data available for later years. The figures and tables below summarize the distribution of our data.

4 Methodological issues

In this section, we give an overview over the methods employed for the analysis of the German spelling reform.

- **Data acquistion**. One of the first critical questions is how to obtain lists of pairs of tokens affected by the spelling reform, i.e. lists of word pairs where each pair represents a pre-reform/post-reform token, e.g. *Spaghetti/Spagetti*. Broadly

[2] And thus including texts from countries other than Germany, e.g. Austria and Switzerland.

Year	Newspaper																															Sum
	1				5					10					15					20					25					30			
1985								x																								1
1986								x																	x							2
1987								x																	x							2
1988								x																	x							2
1989																									x							1
1990																		x							x							2
1991															x	x			x						x							4
1992															x	x			x						x							4
1993				x		x									x	x			x			x		x								7
1994				x											x	x	x	x				x		x					x			8
1995				x		x								x	x	x	x	x						x					x			9
1996				x		x					xx			x	x	x	x	x			x			x			x		x			13
1997		xx		x		xx					xxxx			x	x	x	x	xx			x			xx			x		x			19
1998		xx		x		x					xxxx			x	x	x	x	xx			x			xx					x			17
1999		xx				xx					xxxx			x	x	x	x	xx			x			xx					x			17
2000		xx				x					xx			x	x	x	x	x			x			xx					x			14
2001		xx				x								x			x				x			x					x			8
2002		x												x			xx				x			x					x			7
2003		x				x								x			xx				x			x					x			8
2004		x												x			xx				x			x					x			7
2005		xx				x			x					x			xx				x			xx					x			11
2006		xx		x					x					x			xx				x			xx	x				x			12
2007		xxxx		x			x	x						x			xxx				xx			xx	x				x			17
2008		xxxx		x			x	x						x			xxx				xx			xx	x				x			17
2009		x	xx	x			x	x						x			xxx				xx			xx	x				x			16

Table 1: Distribution of newspaper corpora over years, where newspapers are abbreviated with numbers (1 to 31). The names of the included newspapers are listed on Institut für Deutsche Sprache (2010).

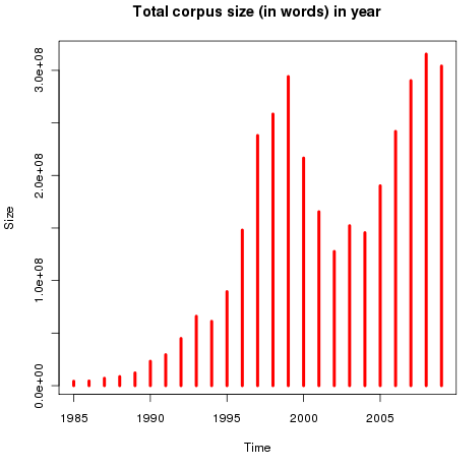

Figure 1: Total per-year-size of corpus (sum over newspapers) measured in number of words for the time slice under consideration.

speaking, there are two alternatives for arriving at such word pair lists. First, we can acquire them '*manually*', e.g. by consulting reference manuals dealing with the spelling reform. Secondly, we could extract them *automatically* from the data by investigating its properties. One possibility for doing this would be to compare two large bodies of data — a pre- and a post-reform portion — by means of log-likelihood ratios or the like (e.g. Dunning (1993)) and thus find "outlier observations" — those stemming from one part of the corpus that are unusually frequent or infrequent with regard to the respective other part — in any of the two subdivisions. These outliers would be good candidates for spelling reform variants provided the corpus subdivision respects the timing of the implementation of the spelling reform.[3]

The first approach has the advantage that in this way only words veritably affected by the reform are considered, but has the disadvantage of possibly overseeing important data points. Moreover, in this way it is usually not possible to include words on which the reform had no direct but only indirect impact (such as 'false analogies', etc.). While the second approach is able to overcome

[3] A more intriguing approach to detecting spelling reform variants is the following. A word form x is (most probably) a spelling reform variant of a word form y if (1) x and y are **formally** similar, where we define this similarity by word edit distance or any other string metric (cf. Cohen et al. (2003)), (2) x and y are **semantically** similar (cf. Jiang and Conrath (1996), Eger and Sejane (2010)), (3) x and y are **competitors**, where this competition would be defined via the words' time series behavior.

these problems, it usually implies a lot of time-intensive manual screening of the automatically extracted candidate words. In the sequel, we will make use of both possibilities of data acquistion.

- **Data representation**. Given our interest in the relative diachronic distribution of competing variants, the data objects under investigation in the current study are token pairs of the form *Spaghetti/Spagetti*, with meaning as indicated above. We represent each such pair by a *single* time series (see below) of the form $\left(\frac{y_{t,\text{new}}}{y_{t,\text{old}}+y_{t,\text{new}}}\right)_{t\in\mathcal{T}}$, where $y_{t,\text{new}}$ stands for the frequency count observation of the post-reform linguistic item at time point t — we describe below how this value is computed — and $y_{t,\text{old}}$ similarly stands for the pre-reform count, $t \in \mathcal{T} \equiv \{1985, 1986, \ldots, 2009\}$.[4]

- **Frequency computations**. In the current study, we include all newspaper data sets available in the DeReKo archive satisfying the time restrictions specified in the preceding section. In other words, we do not explicitly account for a data distribution balanced with respect to geographical, regional or other parameters. Thus, we include, for example, data sets of Swiss and Austrian origin, for which there might exist peculiar orthographical idiosyncrasies; for instance, the letter *ß* has not been part of the Swiss alphabet, neither before nor after the spelling reform. However, we try to correct for "outlier" observations by excluding all data points below or above some fraction of the 'average' observation:

  - For a given year $t \in \{1985, \ldots, 2009\}$ and a given token form z, let $z_{1,t}, \ldots, z_{n,t}$ be all normalized (or, relative)[5] observed frequency counts of z for the n newspaper corpora available for year t, and let m_t and s_t denote the mean value and the standard deviation of this sequence of observations, respectively. Then we exclude frequency count observation i, $1 \leq i \leq n$, if and only if

    $$\left|z_{i,t} - m_t\right| \geq k \cdot s_t \tag{1}$$

    for some a priorily fixed $k \in \mathbb{R}^+$ (e.g. $k = 2$). The final 'corrected' frequency observation for word form z in year t is then the average over the remaining observations. The effect this has is exemplified in Figure 2.

    The frequency adjustments (cf. Gries (2008), Gries (2010)) we make here are motivated by the fact that we are interested in general language behavior (as opposed to, say, the language behavior in a specific newspaper organ) and hence want to discard observations that are too strongly deviant from that average.

[4] The idea behind the illustrated representation is that we ask what part of the total frequency mass of two variant forms in a given year is attributable to the post-reform variant.

[5] Of course, we have to normalize here by the size of the respective corpora.

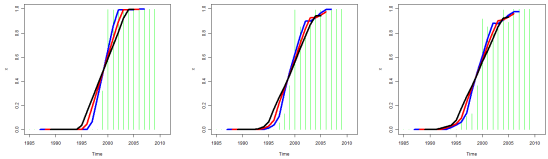

Figure 2: Effects of choosing k in Equation (1) equal to $2, 3, \infty$ (from left to right) on the time series representing the word pair *daß/dass*. Including outlier observations (larger k) in this case entails an illustration of a more 'noisy' time series. Note: In these and all subsequent time series graphics we also depict three moving average trend lines, corresponding to history sizes of 2, 3 and 4.

- **Data analysis**. In chronicling the effects of the German spelling reform on German orthographic language use, it might be useful to generalize over individual linguistic items, e.g. individual token pairs, and thus obtain 'classes' of items behaving similarly — say, all word pairs whose reformed variant becomes dominant after some time period, etc. Such an analysis could identify 'major trends' as an addition to the individual case studies. While it could potentially be conducted 'by hand', in case of hundreds or thousands of data points, a manual inspection would be extremely time consuming, if feasible at all. Therefore, we rely on computational and statistical aids, where necessary. These aim at either (1) explaining the individual time series representing token pairs and/or making them more accessible, or (2) automatically finding classes of word pairs with similarities such as analogical evolution over time, etc.
  - **Time Series analysis**. A time series (e.g. Shumway and Stoffer (2006)) is a sequence of data points related in some way. For example, an AR(1) process is a sequence $\{Y_1, Y_2, \dots\}$ obeying the rule $Y_t = \alpha_1 Y_{t-1} + \epsilon_t$, where α_1, $|\alpha_1| < 1$, is a real parameter and ϵ_t is white noise. The goal of time series analysis is to find an appropriate model for a given observed sequence of data points and thus to be able to make adequate statements about this data. In the given setting of the German spelling reform, modeling our data as time series is quite a natural conceptualization. Yet, despite our belief that such an analysis is extremely useful for understanding our data and thus even for making predictions about future realizations of this data, we will in the current study renounce on intricate statistical time series techniques, mainly for the sake of clarity and simplicity of the given examinations. In the graphical analyses, however, we will include trend lines assisting in the visual interpretation of the given time series graphs.[6]

[6] However, we want to emphasize that future work in this area should assign an appropriate amount of time to more sophisticated time series techniques. Too often in (applied) linguistics a curve or other data is taken at face value, without appropriate statistical tests, etc. Future work on the German spelling reform should hence address problems of stationarity/non-stationarity, integratedness of order d, co-integration, etc., of the time series under analysis.

- **Clustering**. We perform k-means clustering on the time series representing token pairs in order to find variant pairs with similar diachronic behavior.
- **Graphical analysis**. As a second aid to finding token pairs with analogical behavior we will employ 'manual clustering' on the basis of graphical investigation: first, we represent our time series in a lower dimensional space (usually $\mathbb{R}^2$; note that the original data is in $\mathbb{R}^{25}$) by means of principal component analysis (PCA) (e.g. Gentle (2009)) and then assess the results by inspection.

In the following, we summarize our metholodigical approach by putting the steps involved in sequential order.

1. Let a list of token pairs be given, where each pair is affected in some way by the German spelling reform of 1996/2004/2006. This list could have been automatically derived from the DeReKo archive or manually construed. From this list, token pairs that do not fulfill certain frequency restrictions are removed because very low frequency values would considerably limit the reliability of the results.

2. Next, we determine for each of the two variants comprising each token pair average normalized frequency values for all years between 1985 and 2009, excluding outlier frequency observations. The basis of these frequency counts are all newspaper corpora from the DeReKo archive.

3. On the basis of the frequency counts of both variants, we represent each token pair in the list by a single time series, where the frequency of the 'new' form is set in relation to the total frequency mass of the token pair.

4. Finally, we analyze our data. Individual time series are examined by making use of time series analysis techniques (particularly, trend lines) and we try to find classes of token pairs with similar developments using both clustering and graphical analyses.

The above procedure will be applied to all categories affected by the German spelling reform and listed in Section 2.

5 Results

5.1 ASL

Here, we make use of a list of 237 word pairs partially taken from Güthert (2006). We discard word pairs that are too infrequent and perform k-means clustering on the remaining time series for different values of k — where k denotes the number of disjoint 'classes' into which the time series representing token pairs fall —, and, using SSE and silhouette coefficients (cf. Tan et al. (2010)), obtain values of k between 3 and 5 as most reasonable. However, as should be clear from Figure 3, the transition between

classes is blurred here, and exactly how many there are or should be is certainly open to debate. If we consider the value $k = 3$, we can discern that the algorithm has detected three groups of heterogenous developments.

- Words legimitated by the reform that were the dominant variant even before the reform; e.g. *Telefon, Elefant, Babys, Mikrofon, Cleverness, Foto, Porträt*, see also Figure 4, left graph.

- Words that dramatically gained from the reform and that abruptly overwhelmed their competing word form. These include almost all *-ss-* forms, but also words with triple consonants like *Stilllegung*, *Verschlusssache*; many words formerly containing *-ph-* like *Biografie*; the derivations of *-enz* and *-anz* like *Potenzial*, see also Figure 4, right graph.

- Word forms that could not profit from the reform and were not accepted. These include almost all facultative writings of foreign words, e.g. *Portmonee*, *Panter*, *Spagetti*, *Jogurt*, *Ketschup*, etc.; but also words like *aufwänden* and *aufwändig*, see also Figure 5.

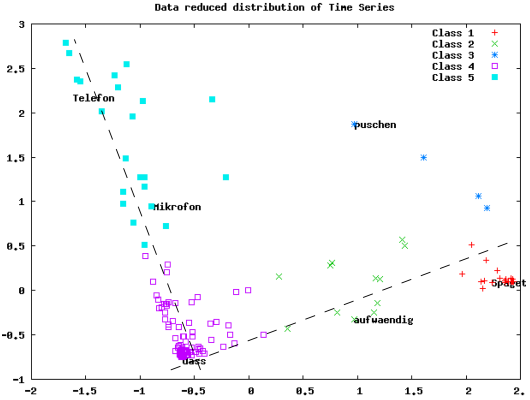

Figure 3: ASL: Representing time series of dimension 25 in a two-dimensional space by means of PCA. Along the new axes determined by PCA (dotted lines) pre-reform establishment of the new variant decreases from top to bottom (along the "new y-axis" going through *Telefon* and *dass*); e.g. while the form *Telefon* was well-established even before the reform, the form *dass* was virtually non-existant (cf. Figure 4). From left to right, the degree of post-reform establishment reduces; i.e. while the form *dass* is now almost completely accepted, the form *Spagetti* is very rare even today (cf. Figure 5). In this figure, using different colors, we also depict five classes found by the k-means algorithm.

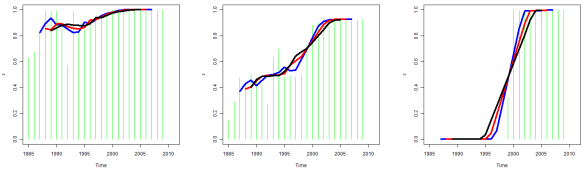

Figure 4: ASL: Words profiting from the spelling reform, ordered by pre-reform acceptance. From left to right: *Telephon/Telefon, Mikrophon/Mikrofon, daß/dass*.

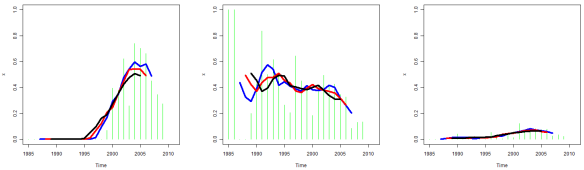

Figure 5: ASL: Words not profiting from the spelling reform, ordered by post-reform acceptance. From left to right: *aufwendig/aufwändig, pushen/puschen, Spaghetti/Spagetti*. The reformed word *puschen* is an exception in that its pre-reform frequency seems to be higher than its post-reform frequency.

5.2 W12

W12 is more difficult to analyze than ASL. First, while the latter class is more or less closed or at least 'easily' representable by a few prominent members, the former class is principally unbounded (*weiter fahren*, *weiter laufen*, *weiter rennen*, ...). Moreover, W12 usually correlates (or correlated) with a meaning differentiation, e.g. *er hat wieder gewählt* (he has voted again) vs. *er wurde wiedergewählt* (he was re-elected), so that it is generally true that *both* of two variant spellings were present both before and after the spelling reform. In order to tackle the first problem, instead of relying on word lists generated by linguistic intuition, we extracted such a list in a data driven way from our corpora in the manner described in Section 4.

This resulted in a list of several thousand entries of token pairs of the form *zusammen sein* vs. *zusammensein* that was in part manually inspected. We then applied the same PCA and clustering analysis as before to the residuary few hundred word pairs, see Figure 6. In the case of W12, we find the following classes of time series distributions:

- Forms whose writing as one word was clearly dominant both before and after the onset of the spelling reform. This includes almost all words containing the prefixes *zusammen-*, *entgegen-*, *fest-*, *mit-*, *weiter-*, and *wieder-*, cf. Figure 7. We note two things about the words in this class:

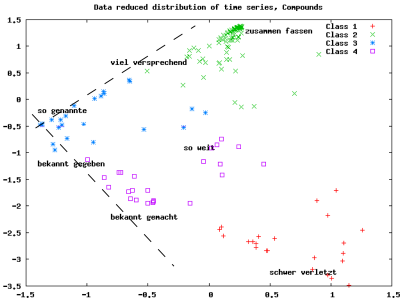

Figure 6: W12: Two-dimensional representation of time series. Along the new axes, acceptance from 1999 to 2006 is increasing from top to bottom (along the axis going through *viel versprechend* and *so genannte*). From left to right, post-reform (particularly, from 2006 onwards) acceptance is increasing.

(i) With the exception of only few words like *gut geheißen*, *schwer wiegend*, etc. the words in this class were usually not subjected to changes prescribed by the spelling reform, which seems to be in accordance with their distributional development.

(ii) Even though in general the spelling reform had not ordered any change, there was usually a slight increase in the writing as two words of the word forms pertaining to this class, most probably as a reflex to the reform ('false analogy').

- Word forms whose writing as two words enormously increased in the beginning years of the reform (usually in 1999) and whose writing as two words decreased again in 2006, when many of the prescriptions of the reform were made optional, cf. Figure 8. In terms of the decrease after 2006, we find tokens here that have a (i) very large (e.g. *lahm gelegt*, *so genannte*, *offen legen*, etc.) (ii) a mediocre (e.g. *bekannt gegeben*, *schwer kranken*, *schief gehen*, etc.), and (iii) a very slight decrease after 2006 (e.g. *übrig gebliebenen*, *nahe gelegenen*, *gefangen genommen*, etc.)

- Word forms whose writing as two words was predominant before the reform, (slightly) increased with the beginning of the reform and has stabilized afterwards, e.g. *schwer verletzt*, *ernst nehmen*, *ernst genommen*, etc., cf. Figure 9.

5.3 Hyphenation

Here, we analyze two different classes of token forms subjected to modification by the spelling reform.

First, we examine numbers suffixed by the forms *er*, *ers*, *fach*, *jährig*, *köpfig*, *mal*, *minütig*, *prozentig*, *seitig*, *stel*, *stellig*, *sten*, *stöckig*, *stündig*, *tägig*, *teilig*, for which, in

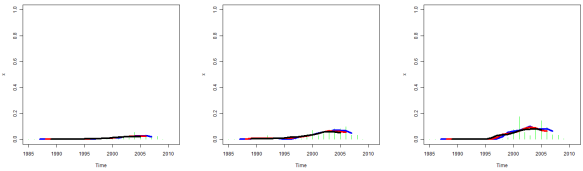

Figure 7: W12: Word forms whose writing as two words was not generally accepted. From left to right: *mitgerechnet/mit gerechnet, zusammenfinden/zusammen finden, schwerwiegend/schwer wiegend.*

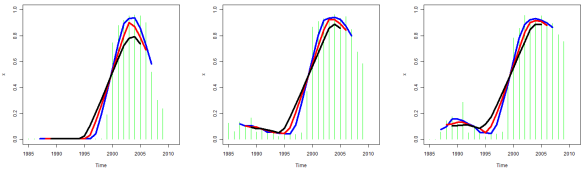

Figure 8: W12: Word forms whose writing as two words increased first and then decreased after 2006. Sorted by dimishing decrease. From left to right: *sogenannte/so genannte, bekanntgegeben/bekannt gegeben, übriggebliebenen/übrig gebliebenen.*

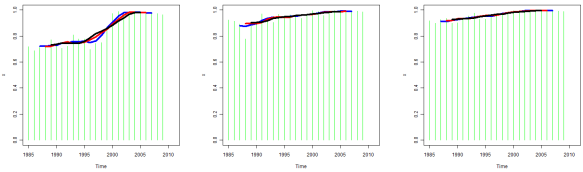

Figure 9: W12: Word forms whose writing as two words has been accepted, sorted by increasing pre-reform establishment. From left to right: *ernstgenommen/ernst genommen, ernstnehmen/ernst nehmen, schwerverletzt/schwer verletzt.*

part, the spelling reform had prescribed the spelling with a hyphen, e.g. 90-*prozentig*, 5-*seitige*, etc. Contrary to our usual procedure, we do not contrast individual token forms here but rather sets of tokens starting with a number, a hyphenation or not, and finally one of the above strings together with possibly other letters — usually inflection markers — like -*e*, -*er*, -*es*, etc.[7]

The results here (cf. Figure 10) clearly indicate the acceptance of the newly prescribed/recommended hyphenation. Whenever the spelling reform decreed its use (e.g.

[7] To put it more algebraically, we contrast here, for example, the sets [0-9]$^+$jährig(e|er|es) / [0-9]$^+$-jährig(e|er|es), etc.

jährig, köpfig, minütig, prozentig, seitig, stöckig, stündig, tägig, teilig), it was indeed frequently and increasingly employed (right graph). For related forms for which the reform did not prescribe its use (*er*, *sten*), we find the already observed 'false analogies' (left graph). In the case of the optional employment of the hyphen in connection with the syllable *fach*, there seems to be a preference for hyphenation, too (middle graph).

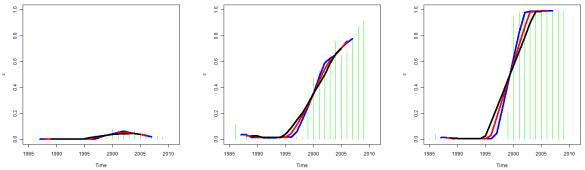

Figure 10: Hyphenation: Different developments for numbers suffixed by forms discussed in the text. From left to right: *er/-er, fach/-fach, jährig/-jährig.*

Secondly, we automatically extracted a list of frequently hyphenated words from DeReKo in the manner described in Section 4, data acquistion. Combining these with their unhyphenated competitors, we generated the usual time series relating two variants of a word form. In this case, however, we cannot find clear trends shared by large majorities of tokens, which might be attributable to the fact that most hyphenation rules prescribed by the spelling reform were optional.[8] Still, we give some tentative judgments based on developments that seem to be discernible from the data.

- Anglicism compounds frequently used in the German language (*Happyend, Callcenter, Talkshow, Internetnutzer, Midlifecrisis*, etc.) seem to 'lose' their hyphenation (cf. Figure 11, first graph), which is in accordance with the recommendations of the reform. The same development seems to be true for combinations with e.g. *Euro-*, for example *Euroraum, Eurozone*.

- Other words like *Tennis-Profi, Bundesliga-Spiel, Co-Trainer, Apartheid-Regime, Nazi-Diktatur, schwarz-weiß, Eishockey-Liga* seem to dsiplay a rather clear trend of a more frequently hyphenated use (cf. Figure 11, last two graphs). However, for many other words it is hard to detect any effect of the spelling reform with regard to hyphenation. Often, changes — if there are any at all — seem to be very slow and also gradual, with no clear breaks at time points relevant for the spelling reform (e.g. 1999, 2004, 2006, etc.).[9]

[8] And possibly to the fact that hyphenation is a rather infrequent phenomenon in the German language anyway, on which not so much emphasis is laid.

[9] On average, hyphenation use seems to have slightly increased, however, in the German language from about 1% to about 1.11% of the tokens. This impression was supported by a Mann-Kendall test for monotonic increase of hyphenation use at the 5% level.

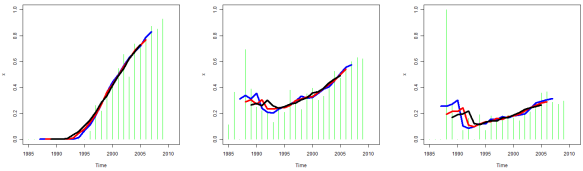

Figure 11: Hyphenation: Left: the general distribution with regard to hyphenation use for anglicism compounds is a steady increase of the variant without hyphen (here *Internet-Nutzer/Internetnutzer*). Middle: Clear trend for increased hyphenation use in *schwarzweiß/schwarz-weiß*. Right: gradual trend for *Tennisprofi/Tennis-Profi*.

5.4 LUC

Our analysis will focus here on three subjects. First, we discuss the spelling of nominalizations like *der Einzelne*, *aufs Beste*, *auf Deutsch*, *des Weiteren*, *in Bezug auf*, *im Folgenden*, *im Allgemeinen*, etc. for which the spelling reform prescribed capitalization of the nominalized part. Secondly, the spelling of adjectival doublets in connection with the verbs *tun*, *gehen*, *geben* and *haben* such as *Pleite gehen*, *Leid tun*, *Recht haben* will be of concern for which, likewise, the reform prescribed capitalization. Finally, we examine the spelling of combinations of adjectives and nouns with proper name character such as *schwarzer Peter*, *schwarzes Brett*, *erste Hilfe*, *heiliger Abend*, etc. where the reform introduced the general rule of using small letters for the adjective part. In all three cases, we rely on word pair lists (of approximately 10-50 instances each) obtained from linguistic reference manuals (e.g. Güthert (2006), Korrekturservice im Internet (2010), etc.).

Concerning the spelling of nominalizations, there is a very clear tendency towards acceptance of the capitalized variants, e.g. *in Bezug auf*, *aufs Beste*, *auf Deutsch*, *im Allgemeinen*, *im Voraus*, *der Einzelne*, etc., cf. Figure 12, left graph. Also, combinations of *gestern, heute, morgen* and times of the day like *gestern Abend*, *heute Mittag* seem to be very well accepted in their capitalized variants. Of particular interest in this connection are combinations like *von Weitem*, *von Neuem*, *von Nahem*, *seit Langem*, *seit Kurzem*, etc. whose capitalized variants were only introduced in the last stage of the reform in 2006. One sees that even here, despite a lack of official regulation and prior to it, capitalization has become slowly more prominent (right graph).

On the other hand, for adjectival doublets it seems that, after 2006, prereform lower case variants are gaining grounds again, see Figure 13. Finally, the situation seems to be still different for combinations of adjectives and nouns with proper name character, where the spelling reform seemed to have little or no success in eliciting alteration, e.g. in establishing predominant use of lower case letters, see Figure 14.

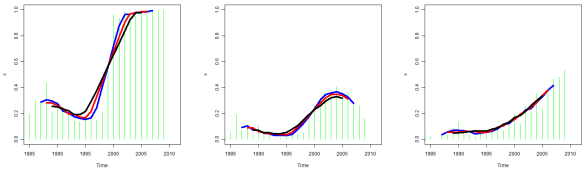

Figure 12: LUC: Left: Capitalization of selected nominalizations like *in bezug auf/in Bezug auf* (shown), *im voraus/im Voraus* has been well accepted. Middle: Decreasing tendencies after 2006 as in *im folgenden/im Folgenden* (shown) or *des weiteren/des Weiteren* seem to be exceptions. Right: Even prior to regulation, there was a trend towards capitalization of further nominalizations as in *von neuem/von Neuem* (shown), *bei weitem/bei Weitem*, *seit langem/seit Langem*, etc.

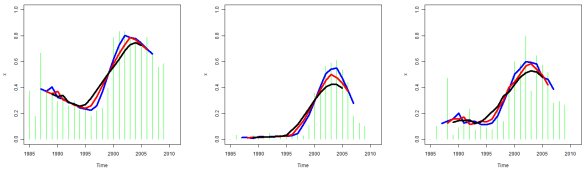

Figure 13: LUC: After 2006, adjectival doublets tend to be spelled with lower case letters again. From left to right: *recht geben/Recht geben, leid tun/Leid tun, pleite gehen/Pleite gehen*.

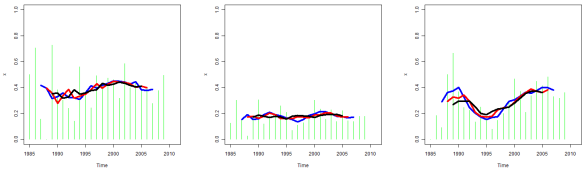

Figure 14: LUC: For combinations of adjectives and nouns with proper name character, the reform did not seem to have noticeable effects. From left to right: *Schwarzen Brett/schwarzen Brett, Erste Hilfe/erste Hilfe, Schwarze Peter/schwarze Peter*.

5.5 Punctuation, end-of-line word separation

For both of these categories we have not conducted corpus based analyses. While punctuation is not easily integrable in our current framework, end-of-line separation of words is usually not performed manually but by means of computer aids and neither is it the goal of the current survey to investigate the quality of these tools nor their particular functionality.

6 Discussion

A very general outcome of the analysis of the German spelling reform conducted in this paper is that the first effects of the reform on actual language usage were palpable in 1999, three years after the official start of the reform. Another frequently observed pattern in the data is the decline of the reformed spellings after 2006, when many reformations were made optional and pre-reform spellings were permitted again. This is however not universally valid for all tokens affected by the reform and we must distinguish here between the different categories on which the reform exerted its influence:

- For ALS, we note that many changes were accepted beyond 2006 by the language community; for example, *-ss-* instead of *-ß-*, triple consonants, *-f-* for *-ph-* ,*-ys* as plural of English nouns ending in *-y*, *-z-* for *-t-* for derivations of nouns ending in *-enz* or *-anz*. On the other hand, particularly optional variant spellings of foreign words such as *Jogurt*, *Panter*, etc. were not accepted at all, while further individual reformations seem to be accepted on a case-to-case basis. For instance, whereas *Quäntchen* has come to dominate over *Quentchen*, the pair *aufwendig/aufwändig* displays the typical pattern discussed above — the reformed variant is strongly decreasing after 2006.

- For the category W12, we find that, at large, the simplification rule designed by the reformers, which had decreed the writing as two words as the standard case, was rather not accepted by the language community. For most cases we see here instead that the writing as two words is declining from 2006 onwards. However, we also note that — up to 2009 — the share of the reform variant has usually not fallen to its pre-reform level. Moreover, the degree of decline of this variant also depends upon the specific word at hand and may require a single case analysis. For example, for the word form *sogenannte*, one could argue that due to the existence of the abbreviation *sog.* a spelling as two words should naturally vanish again.[10]

  Even in this category, however, there are reform 'winners'. Among these are words that, in common pre-reform language usage, used to be frequently spelled as two words anyway (e.g. *ernst genommen*) and the words *zurzeit* and *mithilfe*. The last two are interesting because they form an exception to the general rule of the reform, namely the writing as two words. One might hypothesize that many people were not aware of this last fact, which may have induced a false belief about the respective status of the pre- and postreform variant. The word forms' relative increase even after 2006 could thus possibly be interpreted as a general reflex against the reform, which has been criticized all throughout its implementation (e.g. Rechtschreibung und "Rechtschreibreform" (2010)).

[10] As a more general rule, one could argue that the more semantic difference there is between writing as one and as two words, the more unlikely is it that writing as two words will be commonly accepted. Although our data seems to support this hypothesis (e.g. *übrig geblieben*, *schwer krank* are accepted, *allein erziehende*, *allein stehende* are not), an adequate analysis would be beyond the scope of the methods employed in this paper and is therefore not undertaken.

- The demands of the reform with respect to hyphenation have largely been met. In compounds, numbers are more and more frequently being separated by a hyphen and anglicism compounds are increasingly 'losing' their hyphen. Finally, the degree of hyphenation in the German language seems to have increased since the beginning of the reform, which is in accordance with the reformers' more generous admission of hyphenation usage.

- For the category LUC, we find that capitalization of selected nominalizations has been very much accepted while for adjectival doublets a decrease of capitalization (hence decrease of the reform variant) from 2006 onwards is discernible. The reform did not seem to have much impact on the capitalization of combinations of adjectives and nouns with proper name character.

Before concluding in the next section, we must shortly touch on two further aspects. First, some of the time series representing token pairs seem to have a strange shape in that the reform variant seemed to be more frequent in the late 80s than in the early 90s (cf. Figures 12, 13, etc.). While we do not exactly know the reason for this curvature of the series, they could reflect the idiosyncratic behavior of the few newspaper organs available in our sample for the 1980s. Another explanation could be that the possibly increased application of automatic devices such as spell-checkers over time may have suppressed emergent developments in the German orthographic system.

Secondly, it may be questioned how well a corpus of newspaper articles is suited for addressing problems of language use in a language community. If one is interested in a population parameter (in our case, linguistic behavior of a population of language users) then it is certainly not advisable to consult just a very distinguished subsample of that population. The distribution of lexical tokens in newspaper magazines might be sufficiently different from the distribution in, say, schooling institutions[11], a primary addressee of the spelling reform. In this sense we can only consider our investigation as a (possibly fallible) approximation to the truth.

7 Conclusions and further remarks

In this work, we have presented a generalizable methodological framework for investigating lexical change as induced by the German spelling reform of 1996/2004/2006. This framework includes the acquistion, the representation and the (semi-automatic) analysis of the ('competing') linguistic tokens under scrutiny.

The results of our analysis have shown that some of the entailed changes of the German spelling reform have been *complete* (in the sense that the 'old' variant has been completely substituted by the 'new') while others were only *partial*, and still others were even *reversible*, in the sense that the reform variant is 'dying out' after some period of increase (e.g. the writing as two words of the pre-reform form *sogenannte*).

[11] Particularly, for example, when one thinks of spellings of foreign words such as *Spaghetti/Spagetti*, etc.

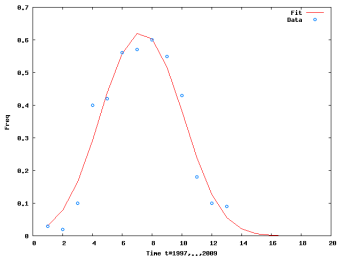

Figure 15: Fitting the logistic representation $\frac{d}{1+ae^{-bt+ct^2}}$ to the data on the pair *leid tun/Leid tun*. Parameter estimates are $a = 280.61$, $b = 1.1894$, $c = 0.0814$, $d = 2.8906$. The fitted line 'predicts' further decreases for the capitalized form, *Leid tun*. The time frame considered is 1997 to 2009, with 'predictions' up to 2013.

Such 'laws of change' have been discovered in quantitative linguistics as governing many language change processes (cf. Altmann (1992)). There, it has also been recognized that these growth developments can be modeled using the logistic representation $y_t = \frac{d}{1+ae^{-bt+ct^2}}$, with appropriate constants $a, b, c, d \in \mathbb{R}$ and where $t \in \mathbb{N}$ is the time index (cf. Best et al. (1990)). Knowing this general structure of language change processes would then be one possibility[12] to project the 'results up-to-now' into the future, i.e. to make *forecasts* (cf. Best (2009)) about developments to come. Figure 15 sketches such a prognosis for the pair *leid tun/Leid tun*. Since information like this can be crucial for 'language engineers' (as language reformers certainly are), this is one place where future work could (and probably should) add on.

8 Acknowledgements

This research was conducted within the project "Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen", financed by the "Bundesministerium für Bildung und Forschung" (BMBF).

[12] Another would of course be to first identify a time series model such as AR(p) as governing the data at hand and then to use corresponding forecasting techniques.

References

Altmann, G. (1992). Piotrowski's law of language change. In *What is Language Synergetics?*, pages 34–35.

Best, K.-H. (2009). Sind prognosen in der linguistik moeglich? In *Typen von Wissen. Begriffliche Unterscheidung und Auspraegungen in der Praxis des Wissenstransfers*, pages 164–175. Lang.

Best, K.-H., Beöthy, E., and Altmann, G. (1990). Ein methodischer beitrag zum piotrowski-gesetz. *Glottometrika*, 12:115–124.

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIW-03)*, pages 73–78.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19/1:61–74.

Eger, S. and Sejane, I. (2010). Computing semantic similarity from bilingual dictionaries. In *Statistical Analysis of Textual Data. Proceedings of the 10th International Conference on statistical analysis of textual data (JADT 2010)*, pages 1217–1225.

Gentle, J. E. (2009). *Computational Statistics*. Springer.

Gries, S. T. (2008). Dispersion and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13/4:403–437.

Gries, S. T. (2010). Useful statistics for corpus linguistics. In *A Mosaic of Corpus Linguistics*, pages 269–291. Lang.

Güthert, K. (2006). Zur neuregelung der deutschen rechtschreibung ab 1. august 2006. *Sprachreport*.

IAO (1992). *Deutsche Rechtschreibung. Vorschläge zu ihrer Neuregelung*. Narr.

Institut für Deutsche Sprache (2010). *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2010-II (Release vom 16.08.2010)*. Institut für Deutsche Sprache, Mannheim. http://www.ids-mannheim.de/kl/projekte/archiv.html.

Jiang, J. and Conrath, D. (1996). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING X)*.

Korrekturservice im Internet (2010). http://www.korrekturen.de/. (retrieved August 2010).

Kupietz, M. and Keibel, H. (2009). The mannheim german reference corpus (dereko) as a basis for empirical linguistic research. *Working Papers in Corpus-based Linguistics and Language Education*, 3:61–76.

Rechtschreibung und "Rechtschreibreform" (2010). http://www.schriftdeutsch.de/. (retrieved September 2010).

Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and Its Applications*. Springer.

Tan, P.-N., Steinbach, M., and Kumar, V. (2010). *Introduction to data mining*. Pearson Addison-Wesley.