Volume 25 - Number 1 - 2010 - ISSN 0175-1336

JLCL

Journal for Language Technology
and Computational Linguistics

Herausgegeben von / Edited by
Lothar Lemnitzer

GSCL Gesellschaft für Sprachtechnologie & Computerlinguistik

Contents

Editorial

Dies ist das erste Heft mit einer neuen Herausgeberschaft. Seit Anfang 2010 bin ich der Herausgeber dieser Verbandszeitschrift der Gesellschaft für Computerlinguistik und Sprachtechnologie.

Damit verbunden ist eine Veränderung in der Publikationspraxis. Es wird keine gedruckten Hefte mehr geben, sondern einzig und allein eine online-Ausgabe. Ich denke, dass diese Publikationspraxis dem Charakter dieser Zeitschrift gerecht wird. Die Mittel, die der Verband durch den Verzicht auf gedruckte Exemplare einspart, werden in die Verbesserung der Webseite (www.jlcl.org) gehen.

Ich möchte mich an dieser Stelle beim Vorstand der GSCL für die Unterstützung bedanken, sie mir jederzeit gewährt wurde. Bedanken möchte ich mich an dieser Stelle auch bei den Autoren der Artikel, die wir für dieses nicht thematisch gebundene Heft versammeln konnten. Auch den Gutachtern gebührt mein Dank für die angenehme Zusammenarbeit.

Schließlich möchte ich mich bei Anna Melzer bedanken, die im Hintergrund für einen reibungslosen Ablauf bei der Entstehung dieses Heftes sorgte.

Am Schluss noch ein Ausblick auf das Jahr 2011: Im Frühjahr wird ein Heft erscheinen, das Beiträge des GSCL-Workshops *Sprachtechnologie und texttechnologische Methoden im E-Learning* bei der KONVENS 2010 in Saarbrücken. Gastherausgeber dieses Heftes sind Maik Stührenberg und Maja Bärenfänger. Im Sommer wird wahrscheinlich wieder ein nicht themengebundenes Heft erscheinen, für das wir uns Beiträge von Ihnen erhoffen. Eine Ausschreibung wird im Frühjahr erscheinen.

Lothar Lemnitzer

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	xx	x		x			xx		x	x								19
1998	xx			x	x					xxxx		x	x	x	xx	x		x			xx			x								17
1999	xx				xx					xxxx		x	x	x	xx	x		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).

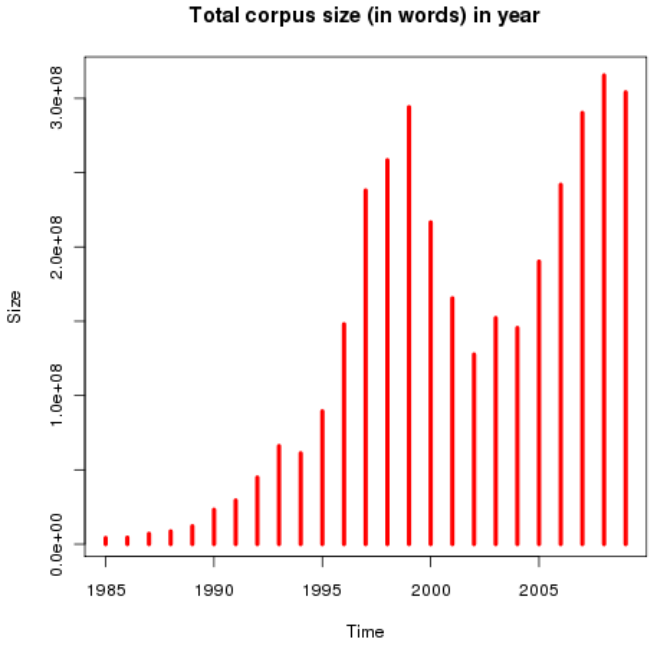

Figure 1: Total per-year-size of corpus (sum over newspapers) measured in number of words for the time slice under consideration.

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.

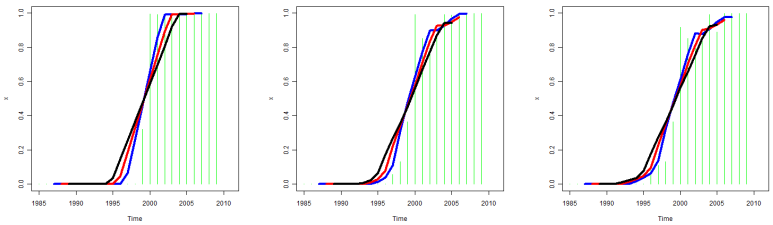

Figure 2: Effects of choosing k in Equation (1) equal to $2, 3, \infty$ (from left to right) on the time series representing the word pair *daß/dass*. Including outlier observations (larger k) in this case entails an illustration of a more 'noisy' time series. Note: In these and all subsequent time series 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, $\left|\alpha_1\right| < 1$, 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.

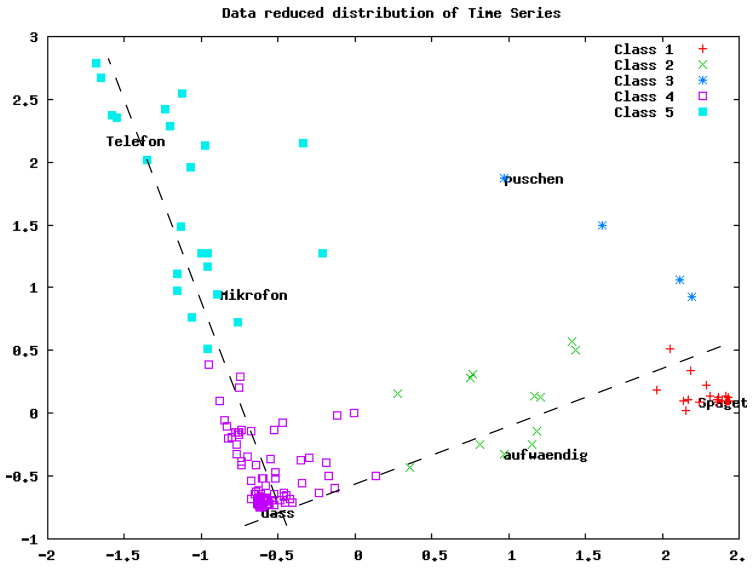

Figure 3: ASL: Representing time series of dimension 25 in a two-dimensional space by means of PCA. Along the new axes determined by PCA (dotted lines) pre-reform establishment of the new variant decreases from top to bottom (along the "new y-axis" going through *Telefon* and *dass*); e.g. while the 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.

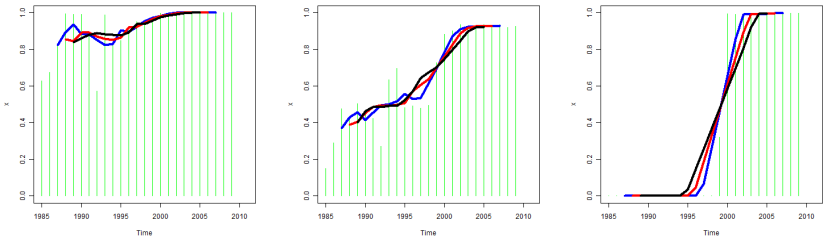

Figure 4: ASL: Words profiting from the spelling reform, ordered by pre-reform acceptance. From left to right: *Telephon/Telefon, Mikrophon/Mikrofon, daß/dass*.

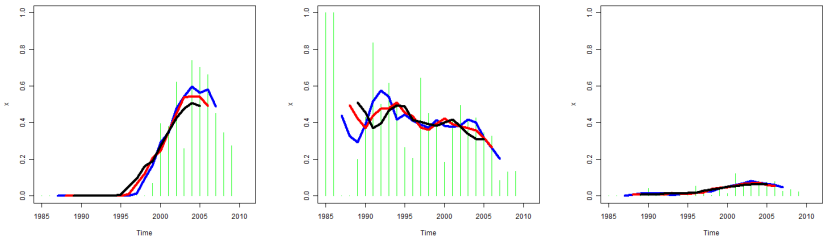

Figure 5: ASL: Words not profiting from the spelling reform, ordered by post-reform acceptance. From left to right: *aufwendig/aufwändig, pushen/puschen, Spaghetti/Spagetti*. The reformed word *puschen* is an exception in that its pre-reform frequency seems to be higher than its post-reform frequency.

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:

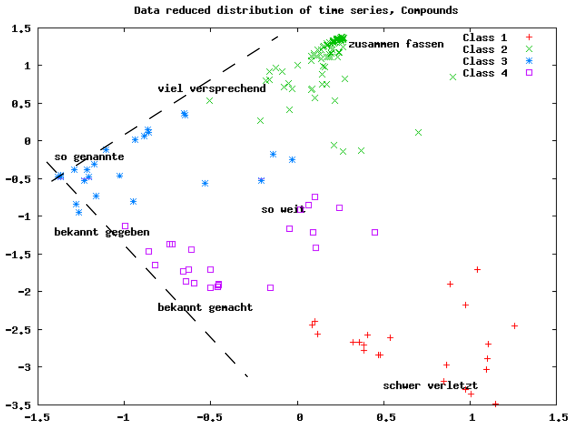

Figure 6: W12: Two-dimensional representation of time series. Along the new axes, acceptance from 1999 to 2006 is increasing from top to bottom (along the axis going through *viel versprechend* and *so genannte*). From left to right, post-reform (particularly, from 2006 onwards) acceptance is increasing.

(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

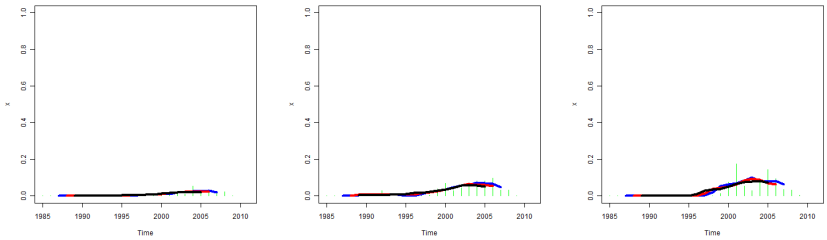

Figure 7: W12: Word forms whose writing as two words was not generally accepted. From left to right: *mitgerechnet/mit gerechnet, zusammenfinden/zusammen finden, schwerwiegend/schwer wiegend*.

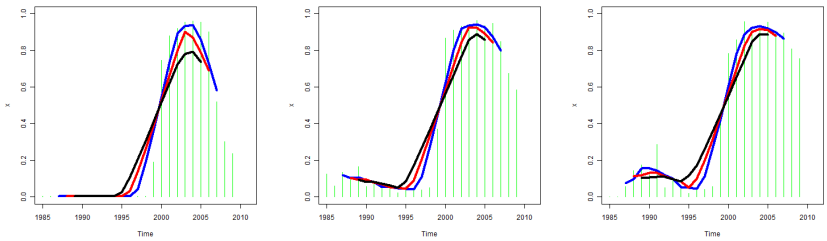

Figure 8: W12: Word forms whose writing as two words increased first and then decreased after 2006. Sorted by dimishing decrease. From left to right: *sogenannte/so genannte, bekanntgegeben/bekannt gegeben, übriggebliebenen/übrig gebliebenen*.

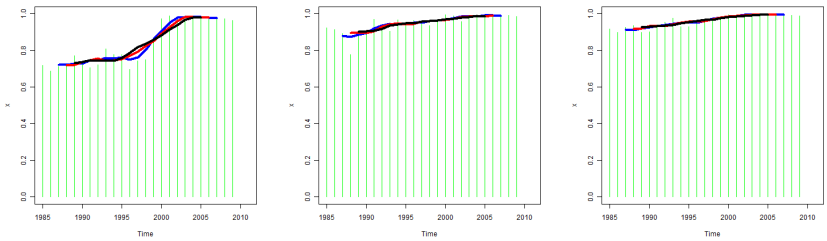

Figure 9: W12: Word forms whose writing as two words has been accepted, sorted by increasing pre-reform establishment. From left to right: *ernstgenommen/ernst genommen, ernstnehmen/ernst nehmen, schwerverletzt/schwer verletzt*.

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).

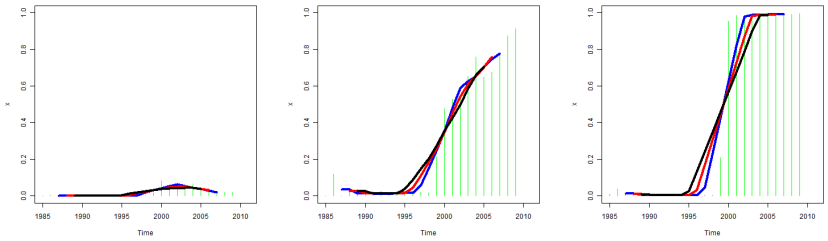

Figure 10: Hyphenation: Different developments for numbers suffixed by forms discussed in the text. From left to right: *er/-er, fach/-fach, jährig/-jährig*.

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.

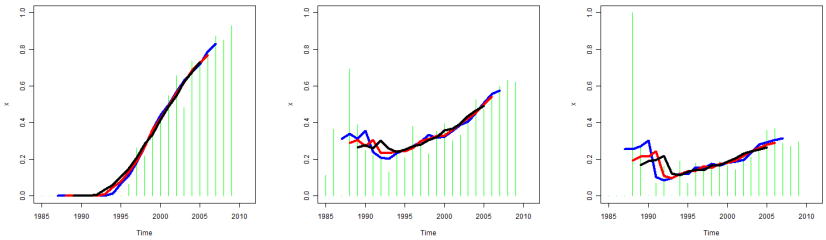

Figure 11: Hyphenation: Left: the general distribution with regard to hyphenation use for anglicism compounds is a steady increase of the variant without hyphen (here *Internet-Nutzer/Internetnutzer*). Middle: Clear trend for increased hyphenation use in *schwarzweiß/schwarz-weiß*. Right: gradual trend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

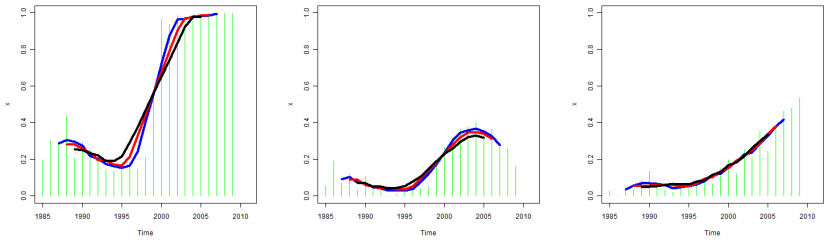

Figure 12: LUC: Left: Capitalization of selected nominalizations like *in bezug auf/in Bezug auf* (shown), *im voraus/im Voraus* has been well accepted. Middle: Decreasing tendencies after 2006 as in *im folgenden/im Folgenden* (shown) or *des weiteren/des Weiteren* seem to be exceptions. Right: Even 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.

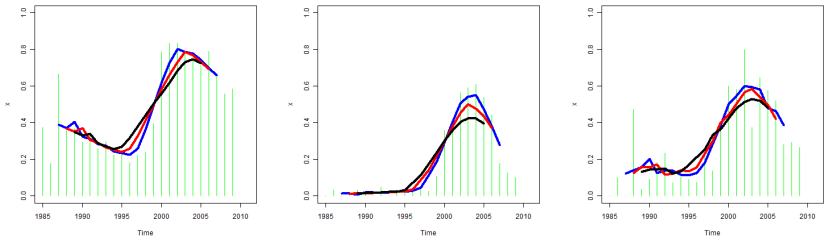

Figure 13: LUC: After 2006, adjectival doublets tend to be spelled with lower case letters again. From left to right: *recht geben/Recht geben, leid tun/Leid tun, pleite gehen/Pleite gehen*.

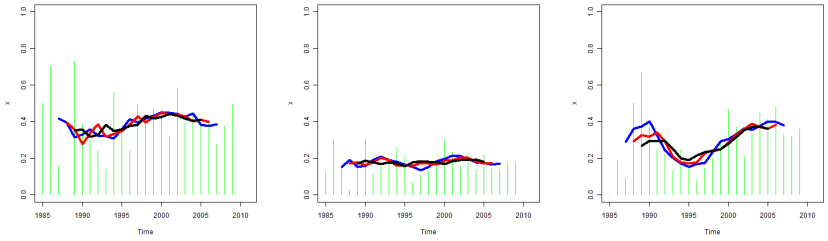

Figure 14: LUC: For combinations of adjectives and nouns with proper name character, the reform did not seem to have noticeable effects. From left to right: *Schwarzen Brett/schwarzen Brett, Erste Hilfe/erste Hilfe, Schwarze Peter/schwarze Peter*.

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.

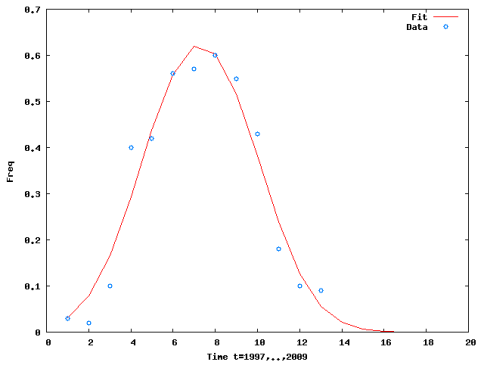

Figure 15: Fitting the logistic representation $\frac{d}{1+ae^{-bt+ct^2}}$ to the data on the pair *leid tun/Leid tun*. Parameter estimates are $a = 280.61$, $b = 1.1894$, $c = 0.0814$, $d = 2.8906$. The fitted line 'predicts' further decreases for the capitalized form, *Leid tun*. The time frame considered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

Bryan Jurish

More Than Words: Using Token Context to Improve Canonicalization of Historical German

1 Introduction

Historical text presents numerous challenges for contemporary natural language processing techniques. In particular, the absence of consistent orthographic conventions in historical text presents difficulties for any system requiring reference to a fixed lexicon accessed by orthographic form, such as information retrieval systems (Sokirko, 2003; Cafarella and Cutting, 2004), part-of-speech taggers (DeRose, 1988; Brill, 1992; Schmid, 1994), simple word stemmers (Lovins, 1968; Porter, 1980), or more sophisticated morphological analyzers (Geyken and Hanneforth, 2006; Zielinski et al., 2009).[1]

Traditional approaches to the problems arising from an attempt to incorporate historical text into such a system rely on the use of additional specialized (often application-specific) lexical resources to explicitly encode known historical variants. Such specialized lexica are not only costly and time-consuming to create, but also – in their simplest form of static finite word lists – necessarily incomplete in the case of a morphologically productive language like German, since a simple finite lexicon cannot account for highly productive morphological processes such as nominal composition (cf. Kempken et al., 2006).

To facilitate the extension of synchronically-oriented natural language processing techniques to historical text while minimizing the need for specialized lexical resources, one may first attempt an automatic *canonicalization* of the input text. Canonicalization approaches (Jurish, 2008, 2010a; Gotscharek et al., 2009a) treat orthographic variation phenomena in historical text as instances of an error-correction problem (Shannon, 1948; Kukich, 1992; Brill and Moore, 2000), seeking to map each (unknown) word of the input text to one or more extant *canonical cognates*: synchronically active types which preserve both the root and morphosyntactic features of the associated historical form(s). To the extent that the canonicalization was successful, application-specific processing can then proceed normally using the returned canonical forms as input, without any need for additional modifications to the application lexicon.

I distinguish between *type-wise* canonicalization techniques which process each input word independently and *token-wise* techniques which make use of the context in which a given instance of a word occurs. In this paper, I present a token-wise canonicalization

[1] While neither information retrieval (IR) systems nor stemmers use a *static* fixed lexicon in the usual sense, the effective lexicon of an IR system is fixed at indexing time as the set of all actually occurring word forms. Similarly, the lexicon of a traditional stemmer has a static portion (hard-coded inflection rules) as well as a dynamic portion (set of stems) determined by the actual input. In both cases, historical spelling variants will be treated as distinct lexemes rather than associated with an equivalent contemporary cognate unless additional measures such as those described here are taken.

method which functions as a disambiguator for sets of hypothesized canonical forms as returned by one or more subordinated type-wise techniques. Section 2 provides a brief review of the type-wise canonicalizers used to generate hypotheses, while section 3 is dedicated to the formal characterization of the disambiguator itself. Section 4 contains a quantitative evaluation of the disambiguator's performance on an information retrieval task over a manually annotated corpus of historical German. Finally, section 5 provides a brief summary and conclusion.

2 Type-wise Conflation

Type-wise conflation techniques are those which process each input word in isolation, independently of its surrounding context. Such a type-wise treatment allows efficient processing of large documents and corpora (since each input type need only be processed once), but disregards potentially useful context information. Formally, a type-wise conflator r is fully specified by a characteristic *conflation relation* $\sim_r$, a binary relation on the set $\mathcal{A}^*$ of all strings over the finite grapheme alphabet $\mathcal{A}$. Prototypically, $\sim_r$ will be a true equivalence relation, inducing a partitioning of the set $\mathcal{A}^*$ of possible word types into equivalence classes or "conflation sets" $[w]_r = \{v \in \mathcal{A}^* : v \sim_r w\}$ induced by some type $w \in \mathcal{A}^*$. Where appropriate, I distinguish between the full conflation set $[w]_r$ containing all strings conflated by r with w and a conflator-specific finite subset $\downarrow[w]_r \subseteq [w]_r$ representing the *canonicalization hypotheses* provided by r for w: the former sets will be used to characterize the retrieval function for r used to define the evaluation measures precision and recall in section 4.2, while the latter will be used in the definition of the token-wise disambiguator in section 3.3. Unless otherwise specified, I assume $\downarrow[w]_r = [w]_r$. In the sequel, I will use the terms "conflation" and "type-wise canonicalization" interchangeably where no ambiguity will result, and the term "conflator" will be used to refer to a specific type-wise canonicalization method.

2.1 String Identity

The simplest of all possible conflators is raw identity of surface strings. The conflation relation $\sim_{\mathrm{id}}$ is in this case nothing more or less than the string identity relation itself:

$$w \sim_{\mathrm{id}} v :\Leftrightarrow w = v \tag{1}$$

String identity is the easiest conflator to implement (no additional programming effort or resources are required) and provides a high degree of precision, "false friends" being limited to historical homographs such as the historical form *wider* when it occurs as a variant of the contemporary form *wieder* ("again") rather than the lexically distinct contemporary homograph *wider* ("against"). Since its coverage is restricted to valid contemporary forms, string identity cannot account for any spelling variation at all, resulting in very poor recall – many relevant types are not retrieved in response to a query in current orthography. Nonetheless, its inclusion as a conflator ensures that the

set of candidate hypotheses $[w]$ for a given input word w is non-empty,[2] and it provides a baseline with respect to which the relative utility of more sophisticated conflators can be evaluated.

As an example, consider the historical form *Abftände*, a variant of the contemporary cognate *Abstände* ("distances"). The conflation set $[\textit{Abftände}]_{\mathrm{id}} = \{\textit{Abftände}\}$ is non-empty, but does not contain the desired contemporary cognate ($\textit{Abstände} \notin [\textit{Abftände}]_{\mathrm{id}}$), so Equation (20) from section 4.2 dictates that no instances of the historical variant *Abftände* will be retrieved via string identity for a query of the contemporary form *Abstände*.

2.2 Transliteration

A slightly less naïve family of conflation methods are those which employ a simple deterministic transliteration function to replace input characters which do not occur in contemporary orthography with extant equivalents. Formally, a transliteration conflator is defined in terms of a character transliteration function $\mathrm{xlit} : \mathcal{A} \to \widetilde{\mathcal{A}}^*$, where $\mathcal{A}$ is as before a "universal" grapheme alphabet (e.g. the set of all Unicode[3] characters) and $\widetilde{\mathcal{A}} \subseteq \mathcal{A}$ is that subset of the universal alphabet allowed by contemporary orthographic conventions. The elementary character transliteration function is extended to a string transliteration function $\mathrm{xlit}^* : \mathcal{A}^* \to \widetilde{\mathcal{A}}^*$ in the usual manner by iteratively applying xlit to each character of the input string in turn (Equation 2), canonicalization hypotheses are limited to the transliterator output (Equation 3), and the characteristic conflation relation $\sim_{\mathrm{xlit}}$ is defined as identity of transliterated strings (Equation 4):

$$\mathrm{xlit}^*(a_1 a_2 \ldots a_n) := \mathrm{xlit}(a_1)\,\mathrm{xlit}(a_2)\ldots\mathrm{xlit}(a_n) \quad (2)$$

$$\downarrow[w]_{\mathrm{xlit}} := \{\mathrm{xlit}^*(w)\} \quad (3)$$

$$w \sim_{\mathrm{xlit}} v :\Leftrightarrow \mathrm{xlit}^*(w) = \mathrm{xlit}^*(v) \quad (4)$$

In the case of historical German, deterministic transliteration is especially useful for its ability to account for typographical phenomena, e.g. by mapping 'ſ' (long 's', as commonly appeared in texts typeset in fraktur) to a conventional round 's', and mapping superscript 'e' to the conventional *Umlaut* diacritic '¨', as in the transliteration *Abftände* $\mapsto$ *Abstände* ("distances"). Given this transliteration, a query for the contemporary form *Abstände* will successfully retrieve all instances of the historical form *Abftände*: $\mathrm{xlit}^*(\textit{Abstände}) = \textit{Abstände} = \mathrm{xlit}^*(\textit{Abftände})$, so $\textit{Abstände} \in [\textit{Abftände}]_{\mathrm{xlit}}$.

The current work makes use of a conservative transliteration function based on the `Text::Unidecode` Perl module.[4] Due to the fact that the underlying character transliteration table is comparatively small and can be implemented as an in-memory

[2] Since $[w]_{\mathrm{id}} = \{w\}$, $[w]_{\mathrm{id}} \subseteq [w]$ implies $w \in [w]$, and thus $[w] \neq \emptyset$. Since the more reliable transliterating conflator described in section 2.2 also ensures a non-empty set of conflation hypotheses, the identity conflator itself was not used to generate hypotheses for the disambiguator in the current experiments.

[3] Unicode Consortium (2011), `http://www.unicode.org/`

[4] `http://search.cpan.org/~sburke/Text-Unidecode-0.04/`

array, transliteration is a very efficient conflation method, with $\mathcal{O}(\text{xlit}) = \mathcal{O}(1)$ and therefore $\mathcal{O}(\text{xlit}^*) = \mathcal{O}(n)$. In terms of expressive power, since xlit is finite, it can be represented by a finite state transducer, and therefore so can its reflexive and transitive closure xlit^*.

Despite its efficiency, and although it outdoes even string identity in terms of its precision, deterministic transliteration suffers from its inability to account for spelling variation phenomena involving extant characters such as the *th/t* and *ey/ei* allographs common in historical German. As an example, consider an instance of the historical form *Theyl* corresponding to the contemporary cognate *Teil* ("part"). Both historical and contemporary forms will be transliterated to themselves, since both strings contain only extant characters, but the historical form will not be retrieved by a query for the contemporary form: $\text{xlit}^*(Teil) = Teil \neq Theyl = \text{xlit}^*(Theyl)$ implies $Teil \not\sim_{\text{xlit}} Theyl$ and therefore $Teil \notin [Theyl]_{\text{xlit}}$.

2.3 Phonetization

A more powerful family of conflation methods is based on the dual intuitions that graphemic forms in historical text were constructed to reflect phonetic forms[5] and that the phonetic system of the target language is diachronically more stable than its graphematic system. Phonetic conflators map each (historical or extant) word $w \in \mathcal{A}^*$ to a unique phonetic form $\text{pho}(w)$ by means of a computable function $\text{pho} : \mathcal{A}^* \to \mathcal{P}^*$,[6] conflating those strings which share a common phonetic form:

$$w \sim_{\text{pho}} v :\Leftrightarrow \text{pho}(w) = \text{pho}(v) \tag{5}$$

Since $[w]_{\text{pho}}$ may be infinite – if for example $\text{pho}(\cdot)$ maps any substring of one or more instances of a single character (e.g. 'a') to a single phone (e.g. [a]) – additional care must be taken to ensure a finite set of canonicalization hypotheses $\downarrow[w]_{\text{pho}}$. A straightforward way to ensure a finite hypothesis set is simply to restrict $[w]_{\text{pho}}$ to some finite set of pre-defined target strings $T \subset \mathcal{A}^*$, setting $\downarrow[w]_{\text{pho}} = \downarrow_T[w]_{\text{pho}} = [w]_{\text{pho}} \cap T$. If pho can be represented as a finite-state transducer M_{pho} and the target lexicon can be represented as a finite-state acceptor A_{Lex}, a more robust alternative is to use a k-best string lookup algorithm such as that described in Jurish (2010b) on the cascade $\mathcal{C}_{\text{pho}}(w) = \text{Id}(w) \circ M_{\text{pho}} \circ M_{\text{pho}}^{-1} \circ A_{\text{Lex}}$, defining $\downarrow[w]_{\text{pho}} = \downarrow_{\mathcal{C},k}[w]_{\text{pho}} = \text{kbest}(k, \mathcal{C}_{\text{pho}}(w))$ for some finite upper bound k on the number of admissible hypotheses, assuming an appropriate weighting scheme on A_{Lex}.

The phonetic conversion module used here was adapted from the phonetization rule-set distributed with the IMS German Festival package (Möhler et al., 2001), a German language module for the Festival text-to-speech system (Black and Taylor, 1997)

[5] Keller (1978) codified this intuition as the imperative "write as you speak" governing historical spelling conventions.

[6] $\mathcal{P}$ is a finite phonetic alphabet.

and compiled as a finite-state transducer (Jurish, 2008).[7] Phonetic conflation offers a substantial improvement in recall over conservative methods such as transliteration or string identity: variation phenomena such as the *th/t* and *ey/ei* allographs mentioned above are correctly captured by the phonetization transducer: pho(*Theyl*) = [taɪl] = pho(*Teil*) which implies $Teil \in [Theyl]_{\text{pho}}$. Unfortunately, these improvements often come at the expense of precision: in particular, many high-frequency types are misconflated by the simplified phonetization rule-set, including **in* ∼ *ihn* ("in" ∼ "him"), **statt* ∼ *Stadt*, ("instead" ∼ "city"), and **wider* ∼ *wieder* ("against" ∼ "again"). While such high-frequency cases might be easily handled in a mature system by a small exception lexicon, the underlying tendency of strict phonetic conflation either to over- or to under-generalize – depending on the granularity of the phonetization function – is likely to remain, expressing itself in information retrieval tasks as reduced precision or reduced recall, respectively.

2.4 Rewrite Transduction

Despite its comparatively high recall, the phonetic conflator fails to relate unknown historical forms with any extant equivalent whenever the graphemic variation leads to non-identity of the respective phonetic forms (e.g. pho(*umb*) = [ʔʊmp] ≠ [ʔʊm] = pho(*um*) for the historical variant *umb* of the preposition *um* ("around")), suggesting that recall might be further improved by relaxing the strict identity criterion on the right hand side of Equation (5). Moreover, a fine-grained and appropriately parameterized conflator should be less susceptible to precision errors than an "all-or-nothing" (phonetic) identity condition (Kondrak, 2000, 2002). A technique which fulfills both of the above desiderata is *rewrite transduction*, which can be understood as a generalization of the well-known *string edit distance* (Damerau, 1964; Levenshtein, 1966).

Formally, let $\text{Lex} \subseteq \mathcal{A}^*$ be the (possibly infinite) lexicon of all extant forms encoded as a finite-state acceptor A_{Lex}, and let M_{rw} be a weighted finite-state transducer over a bounded semiring $\mathcal{K}$ which models (potential) diachronic change likelihood as a weighted rational relation. Then define for every input type $w \in \mathcal{A}^*$ the "best" extant equivalent $\text{best}_{\text{rw}}(w)$ as the unique extant type $v \in \text{Lex}$ with minimal edit-distance to the input word:

$$\text{best}_{\text{rw}}(w) = \arg\min_{v \in \mathcal{A}^*} [\![M_{\text{rw}} \circ A_{\text{Lex}}]\!](w, v) \tag{6}$$

Ideally, the image of a word w under best_{rw} will itself be the canonical cognate sought, leading to conflation of all strings which share a common image under best_{rw}:

$$w \sim_{\text{rw}} v :\Leftrightarrow \text{best}_{\text{rw}}(w) = \text{best}_{\text{rw}}(v) \tag{7}$$

[7] In the absence of a language-specific phonetization function, a general-purpose phonetic digest algorithm such as SOUNDEX (Russell, 1918), the *Kölner Phonetik* (Postel, 1969), PHONIX (Gadd, 1988, 1990), or Metaphone (Philips, 1990, 2000) may be employed instead (Robertson and Willett, 1993; Kempken, 2005).

The current experiments were performed using the heuristic rewrite transducer described in Jurish (2010a), compiled from 306 manually constructed two-level rules, while the lexical target acceptor A_{Lex} was extracted from the TAGH morphology transducer (Geyken and Hanneforth, 2006). The native TAGH weights were scaled for compatibility and used to provide a prior cost distribution over target word forms based on their derivational complexity. Best-path lookup was performed using a specialized variant of the well-known *Dijkstra algorithm* (Dijkstra, 1959) as described in Jurish (2010b). Related approaches to historical variant detection include Kempken (2005); Rayson et al. (2005); Ernst-Gerlach and Fuhr (2006); Gotscharek et al. (2009a).

Although this rewrite cascade does indeed improve both precision and recall with respect to the phonetic conflator, these improvements are of comparatively small magnitude, precision in particular remaining well below the level of conservative conflators such as naïve string identity or transliteration, due largely to interference from "false friends" such as the valid contemporary compound *Rockermehl* ("rocker-flour") for the historical variant *Rockermel* of the contemporary form *Rockärmel* ("coat-sleeve") as appearing in Figure 1.

3 Token-wise Disambiguation

In an effort to recover some degree of the precision offered by conservative conflation techniques such as transliteration while still benefiting from the flexibility and improved recall provided by more ambitious techniques such as phonetization or rewrite transduction, I have developed a method for disambiguating type-wise conflation sets which operates on the token level, using sentential context to determine a unique "best" canonical form for each input token. Specifically, the disambiguator employs a Hidden Markov Model (HMM) whose lexical probability matrix is dynamically re-computed for each input sentence from the conflation sets returned by one or more subordinated type-wise conflators, and whose transition probabilities are given by a static word k-gram model of the target language, in this case contemporary German adhering to current orthographic conventions. Similar approaches for traditional spell-checking applications using strictly local context for language modelling have been described by Kernighan et al. (1990); Church and Gale (1991); Brill and Moore (2000); Verberne (2002). Most closely related to the current proposal is the approach of Mays et al. (1991), who use a word trigram model to disambiguate unweighted confusion sets returned by a traditional approximate Damerau-Levenshtein matcher analogous to the rewrite cascade from section 2.4. An example of the proposed disambiguation architecture for the conflators described in section 2 is given in Figure 1.

3.1 Basic Model

Formally, let $\mathcal{W} \subset \widetilde{\mathcal{A}}^*$ be a finite set of known extant words, let $\mathbf{u} \notin \mathcal{W}$ be a designated symbol representing an unknown word, let $S = \langle w_1, \ldots, w_{n_S} \rangle$ be an input sentence of n_S (historical) words with $w_i \in \mathcal{A}^*$ for $1 \leq i \leq n_S$, and let $R = \{r_1, \ldots, r_{n_R}\}$ be a

	Dete	ſammlete	Steyne	im	Rockermel
id	*Dete*	*ſammlete*	*Steyne*	*im*	*Rockermel*
xlit	*Dete*	*sammlete*	*Steyne*	*im*	*Rockermel*
pho	$\emptyset$	$\emptyset$	{*Steine*}	{*im*, *ihm*}	{*Rockärmel*}
rw	*Tete*⟨1⟩	*sammelte*⟨5⟩	*Steine*⟨1⟩	*im*⟨0⟩	*Rockermehl*⟨10⟩
hmm	*Dete*	*sammelte*	*Steine*	*im*	*Rockärmel*

Figure 1: Example of the proposed conflator disambiguation architecture for the input sentence *"Dete ſammlete Steyne im Rockermel"* ("Dete gathered rocks in the coat-sleeve"). Costs assigned by the rewrite transducer appear in angled brackets, and the conflation hypotheses selected by the HMM disambiguator are underlined.

finite set of (opaque) type-wise conflators. Then, the disambiguator HMM is defined in the usual way (Rabiner, 1989; Charniak et al., 1993; Manning and Schütze, 1999) as the 5-tuple $D = \langle \mathcal{Q}, \mathcal{O}_S, \pi, A, B_S \rangle$, where:

1. $\mathcal{Q} = (\mathcal{W} \cup \{\mathbf{u}\}) \times R$ is a finite set of model *states*, where each state $q \in \mathcal{Q}$ is a pair $\langle \tilde{w}_q, r_q \rangle$ composed of an extant word form $\tilde{w}_q$ and a conflator r_q;
2. $\mathcal{O}_S = \bigcup_{i=1}^{n_S} \{w_i\}$ is the set of *observations* for the input sentence S;
3. $\pi : \mathcal{Q} \to [0,1] : q \mapsto p(Q_1 = q)$ is a static probability distribution over $\mathcal{Q}$ representing the model's *initial state probabilities*;
4. $A : \mathcal{Q}^k \to [0,1] : \langle q_1, \ldots, q_k \rangle \mapsto p(Q_i = q_k | Q_{i-k+1} = q_1, \ldots, Q_{i-1} = q_{k-1})$ is a static conditional probability distribution over state k-grams representing the model's *state transition probabilities*; and
5. $B_S : \mathcal{Q} \times \mathcal{O}_S \to [0,1] : \langle q, o \rangle \mapsto p(O = o | Q = q)$ is a dynamic probability distribution over observations conditioned on states representing the model's *lexical probabilities*.

Using the shorthand notation w_i^{i+j} for the string $w_i w_{i+1} \ldots w_{i+j}$, the model D computes sentential probability as the sum of path probabilities over all possible generating state sequences:

$$p(S = w_1^{n_S}) \quad = \sum_{q_1^{n_S} \in \mathcal{Q}^{n_S}} p(S = w_1^{n_S}, Q = q_1^{n_S}) \tag{8}$$

Assuming suitable boundary handling for negative indices, joint path probabilities themselves are computed as:

$$p(S = w_1^{n_S}, Q = q_1^{n_S}) \quad = \prod_{i=1}^{n_S} p(q_i | q_{i-k+1}^{i-1}) p(w_i | q_i) \tag{9}$$

Underlying these equations are the following Markov assumptions:

$$p(q_i|q_1^{i-1}, w_1^{i-1}) = p(q_i|q_{i-k+1}^{i-1}) \quad (10)$$

$$p(w_i|q_1^{i}, w_1^{i-1}) = p(w_i|q_i) \quad (11)$$

Equation (10) asserts that state transition probabilities depend on at most the preceding $k-1$ states. Equation (11) asserts the independence of observed surface forms (historical spellings) from all but the model's current state. Taken together, these assumptions will lead to the use of a k-gram distribution over contemporary word forms to model both syntactic and (local) semantic constraints of the target language as operating on conflator-dependent type-wise canonicalization hypotheses for historical input forms. Crucially, the product of these two component distributions as used in the path probability computation from Equation (9) will allow linguistic context constraints (insofar as they are captured by the k-gram transition probabilities) to override prior type-wise estimates of a conflation's reliability (and vice versa), leading to a disambiguator dependent on both token context and prior estimates of conflation likelihood.

3.2 Transition Probabilities

The finite target lexicon $\mathcal{W}$ can easily be extracted from a corpus of contemporary text. For estimating the static distributions π and A, we first make the following assumptions:

$$p(Q = \langle \tilde{w}_q, r_q \rangle) = p(W = \tilde{w}_q)p(R = r_q) \quad (12)$$

$$p(R = r) = \frac{1}{n_R} \quad (13)$$

Equation (12) asserts the independence of extant forms and conflators, while Equation (13) assumes a uniform distribution over conflators. Given these assumptions, the static state distributions π and A can be estimated as:

$$\pi(q) :\approx p\left(W_1 = \tilde{w}_q\right)/n_R \quad (14)$$

$$A(q_1, \ldots, q_k) :\approx p\left(W_i = \tilde{w}_{q_k} | W_{i-k+1}^{i-1} = \tilde{w}_{q_1} \ldots \tilde{w}_{q_{k-1}}\right)/n_R \quad (15)$$

Equations (14) and (15) are nothing more or less than a word k-gram model over extant forms, scaled by the constant $\frac{1}{n_R}$. One can therefore use standard maximum likelihood techniques to estimate π and A from a corpus of contemporary text (Bahl et al., 1983; Manning and Schütze, 1999).

For the current experiments, a word trigram model ($k = 3$) was trained on the TIGER corpus of contemporary German (Brants et al., 2002). Probabilities for the "unknown" form **u** were computed using the simple smoothing technique of assigning **u** a pseudo-frequency of $\frac{1}{2}$ (Lidstone, 1920; Manning and Schütze, 1999). To account for unseen trigrams, the resulting trigram model was smoothed by linear interpolation of

uni-, bi-, and trigrams (Jelinek and Mercer, 1980, 1985), using the method described by Brants (2000) to estimate the interpolation coefficients.

3.3 Lexical Probabilities

In the absence of a representative corpus of conflator-specific manually annotated training data, simple maximum likelihood techniques cannot be used to estimate the model's lexical probabilities B_S. Instead, lexical probabilities are instantiated as a Maxwell-Boltzmann distribution for a set d_r of conflator-specific distance functions (Jaynes, 1983):

$$B\big(\langle\tilde{w}, r\rangle, w\big) :\approx \frac{b^{\beta d_r(w,\tilde{w})}}{\sum_{r'\in R}\sum_{\tilde{w}'\in\downarrow[w]_{r'}} b^{\beta d_{r'}(w,\tilde{w}')}} \tag{16}$$

Here, $b, \beta \in \mathbb{R}$ are free model parameters with $b \geq 1$ and $\beta \leq 0$. For a conflator $r \in R$, the function $d_r : \mathcal{A}^* \times \mathcal{W} \to \mathbb{R}_+$ is a pseudo-metric used to estimate the reliability of the conflator's association of an input word w with the extant form $\tilde{w}$, and the set $\downarrow[w]_r \subseteq [w]_r \subseteq \mathcal{A}^*$ is a finite set of canonicalization hypotheses provided by r for w, as described in section 2.

It should be explicitly noted that the denominator of the right-hand side of Equation (16) is a sum over all model states (canonicalization hypotheses) $\langle\tilde{w}', r'\rangle$ actually associated with the observation argument w by the type-wise conflation stage, and *not* a sum over observations w' associable with the state argument $\langle\tilde{w}, r\rangle$. This latter sum (if it could be efficiently computed) would adhere to the traditional form $\big(\mathrm{sim}(o,q)/\sum_{o'}\mathrm{sim}(o',q)\big)$ for estimating a probability distribution $p(O|Q)$ over *observations* conditioned on model states such as the HMM lexical probability matrix B_S is defined to represent; whereas the estimator in Equation (16) is of the form $\big(\mathrm{sim}(o,q)/\sum_{q'}\mathrm{sim}(o,q')\big)$, which corresponds more closely to a distribution $p(Q|O)$ over *states* conditioned on observations.[8]

From a practical standpoint, it should be clear that Equation (16) is much more efficient to compute than an estimator summing globally over potential observations, since all the data needed to compute Equation (16) are provided by the type-wise preprocessing of the input sentence S itself, whereas a theoretically pure global estimator would require a whole arsenal of *inverse* conflators as well as a mechanism for restricting their outputs to some tractable set of admissible historical forms, and hence would be of little practical use. From a formal standpoint, I believe that Equation (16) as used in the run-time disambiguator can be shown to be equivalent to a global estimator, provided that the conflator pseudo-metrics d_r are symmetric and the languages of both historical and extant forms have identical and uniform density with respect to the d_r, but a proof of this conjecture is beyond the scope of this paper.

It was noted above in Section 2.3 that for the phonetic conflator in particular, the equivalence class $[w]_{\mathrm{pho}} = \{v \in \mathcal{A}^* : w \sim_{\mathrm{pho}} v\}$ may not be finite. In order to ensure the

[8] See the discussion surrounding Equation 20 in Charniak et al. (1993) for a more detailed look at these two sorts of lexical probability estimator and their effects on HMM part-of-speech taggers.

computational tractability of Equation (16) therefore, the phonetic conflation hypotheses considered were implicitly restricted to the finite set $\mathcal{W}$ of known extant forms used to define the model's states, $\downarrow[w]_{\text{pho}} = \downarrow_{\mathcal{W}}[w]_{\text{pho}} = [w]_{\text{pho}} \cap \mathcal{W}$. Transliterations and rewrite targets which were not also known extant forms were implicitly mapped to the designated symbol **u** for purposes of estimating transition probabilities for previously unseen extant word types.

For the current experiments, the following model parameters were used:

$$\begin{aligned}
b &= 2 \\
\beta &= -1 \\
R &= \{\text{xlit}, \text{pho}, \text{rw}\} \\
d_{\text{xlit}}(w, \tilde{w}) &= 2/|w| && \text{if } \tilde{w} = \text{xlit}^*(w) \\
d_{\text{pho}}(w, \tilde{w}) &= 1/|w| && \text{if } \tilde{w} \in \downarrow[w]_{\text{pho}} \\
d_{\text{rw}}(w, \tilde{w}) &= [\![M_{\text{rw}} \circ A_{\text{Lex}}]\!](w, \tilde{w})/|w| && \text{if } \tilde{w} = \text{best}_{\text{rw}}(w)
\end{aligned}$$

In all other cases, $d_r(w, \tilde{w})$ is undefined and $B(\langle \tilde{w}, r \rangle, w) = 0$. Note that all conflator distance functions are scaled by inverse input word length $\frac{1}{|w|}$, thus expressing an average distance per input character as opposed to an absolute distance for the input word. Defining distance functions in terms of (inverse) word length in this manner captures the intuition that a conflator is less likely to discover a false positive conflation for a longer input word than for a short one; natural language lexica tending to be maximally dense for short (usually closed-class) words.[9] The transliteration and phonetic conflators are constants given input word length, whereas the rewrite conflator makes use of the cost $[\![M_{\text{rw}} \circ A_{\text{Lex}}]\!](w, \tilde{w})$ assigned to the conflation pair by the rewrite cascade itself.

3.4 Runtime Disambiguation

Having defined the disambiguator model D, it can be used it to determine a unique "best" canonical form for each input sentence S by application of the well-known *Viterbi algorithm* (Viterbi, 1967). Formally, the Viterbi algorithm computes the state path with maximal probability for the observed sentence:

$$\textsc{Viterbi}(S, D) = \underset{\langle q_1, \ldots, q_{n_S} \rangle \in \mathcal{Q}^{n_S}}{\arg\max} p(q_1, \ldots, q_{n_S}, S|D) \tag{17}$$

Extracting the disambiguated canonical forms $\hat{S} = \langle \hat{w}_1, \ldots, \hat{w}_{n_S} \rangle \in (\mathcal{A}^*)^{n_S}$ from the state sequence $\hat{Q} = \langle \hat{q}_1, \ldots, \hat{q}_{n_S} \rangle = \textsc{Viterbi}(S, D)$ returned by the Viterbi algorithm is a simple matter of projecting the extant word components of the HMM state structures, taking care to map the designated symbol **u** onto an appropriate output

[9] Despite this tendency of natural languages, the combinatorial properties of concatenative monoids dictate that the number of potential "false friends" grows exponentially with input string length if for example arbitrary substitutions are allowed, suggesting an increased likelihood of false positive conflations for *longer* input words. In this context, note that the use of per-character distances results in higher-entropy probability distributions (Shannon, 1948) for longer input strings, effectively treating the d_r distance estimates as increasingly unreliable as input string length grows.

string. Let witness : $\wp(\mathcal{A}^*) \rightarrow \mathcal{A}^*$ be a choice function over conflation hypotheses,[10] witness$(\downarrow[w]_r) \in \downarrow[w]_r$ for all $w \in \mathcal{A}^*, r \in R$ with $\downarrow[w]_r \neq \emptyset$, and for $1 \leq i \leq n_S$, define:

$$\hat{w}_i := \begin{cases} \text{witness}\left(\downarrow[w]_{r_{\hat{q}_i}}\right) & \text{if } \tilde{w}_{\hat{q}_i} = \mathbf{u} \\ \tilde{w}_{\hat{q}_i} & \text{otherwise} \end{cases} \tag{18}$$

Following the equivalence class notation for type-wise conflators, I write $[w_i]_{\text{hmm},D}$ to denote the singleton set $\{\hat{w}_i\}$ containing the unique canonical form returned by the HMM disambiguator D for an input token w_i in sentential context S, omitting the model subscript D where no ambiguity will result.

3.5 Expressive Power

It was noted in section 2 above that each of the type-wise conflators used in the current approach have representations as (weighted) finite-state transducers (WFSTs). Since the union of WFSTs is itself a WFST, as is the concatenation of WFSTs (Mohri, 2009), the type-wise analysis stage which generates canonicalization hypotheses for the disambiguator can be expressed by an extended rational algebraic expression, assuming specialized functions such as the k-best lookup used by the rewrite transducer are included in the inventory of admissible operations. Hidden Markov Models have been shown to be equivalent to the sub-family of WFSTs called probabilistic finite-state automata (PFSAs) by Vidal et al. (2005). Pereira and Riley (1997) advocate a decomposition of HMM component distributions into dedicated WFSTs which may then be cascaded (composed) to simulate the original HMM for use in speech processing applications. Hanneforth and Würzner (2009) present a technique for creating n-gram language models using only the algebra of weighted rational languages which can in principle be extended to implement the disambiguator's dynamic lexical probability distribution given by Equation (16) as just such a dedicated WFST component. Finally, since the Viterbi algorithm can be applied directly to PFSAs (Vidal et al., 2005) and with minimal adaptation to appropriately weighted WFSTs (Mohri, 2002; Jurish, 2010b), the entire proposed canonicalization architecture does not exceed the expressive power of the weighted rational relations.

4 Evaluation

4.1 Test Corpus

The conflation and disambiguation techniques described above were tested on a manually annotated corpus of historical German drawn from the *Deutsches Textarchiv*.[11] The test corpus was comprised of the full body text from 13 volumes published between 1780 and 1880, and contained 152,776 tokens of 17,417 distinct types in 9,079 sentences,

[10] Since conflation hypothesis sets $\downarrow[w]_r$ are finite, the axiom of choice is not strictly required here.
[11] `http://www.deutschestextarchiv.de`

discounting non-alphabetic types such as punctuation. To assign an extant canonical equivalent to each token of the test corpus, the text of each volume was automatically aligned token-wise with a contemporary edition of the same volume. Automatically discovered non-identity alignment pair types were presented to a human annotator for confirmation. In a second annotation pass, all tokens lacking an identical or manually confirmed alignment target were inspected in context and manually assigned a canonical form. Whenever they were presented to a human annotator, proper names and extinct lexemes were treated as their own canonical forms. In all other cases, equivalence was determined by direct etymological relation of the root in addition to matching morphosyntactic features. Problematic tokens were marked as such and subjected to expert review. Marginalia, front and back matter, speaker and stage directions, and tokenization errors were excluded from the final evaluation corpus.

4.2 Evaluation Measures

The canonicalization methods from sections 2 and 3 were evaluated using the gold-standard test corpus to simulate an information retrieval task Formally, let $C = \{c_1, \ldots, c_{n_C}\}$ be a finite set of canonicalizers, and let $G = \langle g_1, \ldots, g_{n_G} \rangle$ represent the test corpus, where each token g_i is a $(2 + n_C)$-tuple $g_i = \langle w_i, \tilde{w}_i, [w_i]_{c_1}, \ldots, [w_i]_{c_{n_C}} \rangle \in \mathcal{A}^* \times \mathcal{A}^* \times \wp(\mathcal{A}^*)^{n_C}$, for $1 \leq i \leq n_G$. Here, w_i represents the literal token text as appearing in the historical corpus, $\tilde{w}_i$ is its gold-standard canonical cognate, and $[w_i]_{c_j}$ is the set of canonical forms assigned to the token by the canonicalizer c_j, for $1 \leq j \leq n_C$. Let $Q = \bigcup_{i=1}^{n_G} \{\tilde{w}_i\}$ be the set of all canonical cognates represented in the corpus, and define for each canonicalizer $c \in C$ and query string $q \in Q$ the sets $\text{relevant}(q), \text{retrieved}_c(q) \subset \mathbb{N}$ of *relevant* and *retrieved* corpus tokens as:

$$\text{relevant}(q) = \{i \in \mathbb{N} : q = \tilde{w}_i\} \tag{19}$$

$$\text{retrieved}_c(q) = \{i \in \mathbb{N} : q \in [w_i]_c\} \tag{20}$$

Token-wise precision ($\text{pr}_{\text{tok},c}$) and recall ($\text{rc}_{\text{tok},c}$) for the canonicalizer c can then be defined as:

$$\text{pr}_{\text{tok},c} = \frac{\left|\bigcup_{q \in Q} \text{retrieved}_c(q) \cap \text{relevant}(q)\right|}{\left|\bigcup_{q \in Q} \text{retrieved}_c(q)\right|} \tag{21}$$

$$\text{rc}_{\text{tok},c} = \frac{\left|\bigcup_{q \in Q} \text{retrieved}_c(q) \cap \text{relevant}(q)\right|}{\left|\bigcup_{q \in Q} \text{relevant}(q)\right|} \tag{22}$$

Type-wise measures $\text{pr}_{\text{typ},c}$ and $\text{rc}_{\text{typ},c}$ are defined analogously, by mapping the token index sets of Equations (19) and (20) to corpus types before applying Equations (21) and (22). I use the unweighted harmonic precision-recall average F (van Rijsbergen,

	% Types			% Tokens		
c	$\mathrm{pr}_{\mathrm{typ}}$	$\mathrm{rc}_{\mathrm{typ}}$	$\mathrm{F}_{\mathrm{typ}}$	$\mathrm{pr}_{\mathrm{tok}}$	$\mathrm{rc}_{\mathrm{tok}}$	$\mathrm{F}_{\mathrm{tok}}$
id	99.0	59.2	74.1	99.8	79.3	88.4
xlit	**99.1**	89.5	94.1	**99.8**	96.8	98.3
pho	97.1	96.1	96.6	91.4	99.2	95.1
rw	97.6	**96.5**	**97.0**	94.3	**99.3**	96.7
hmm	98.6	95.3	96.9	99.7	99.1	**99.4**

Table 1: Evaluation data for various canonicalization techniques with respect to the *Deutsches Textarchiv* evaluation subset. The maximum value in each column appears in boldface type.

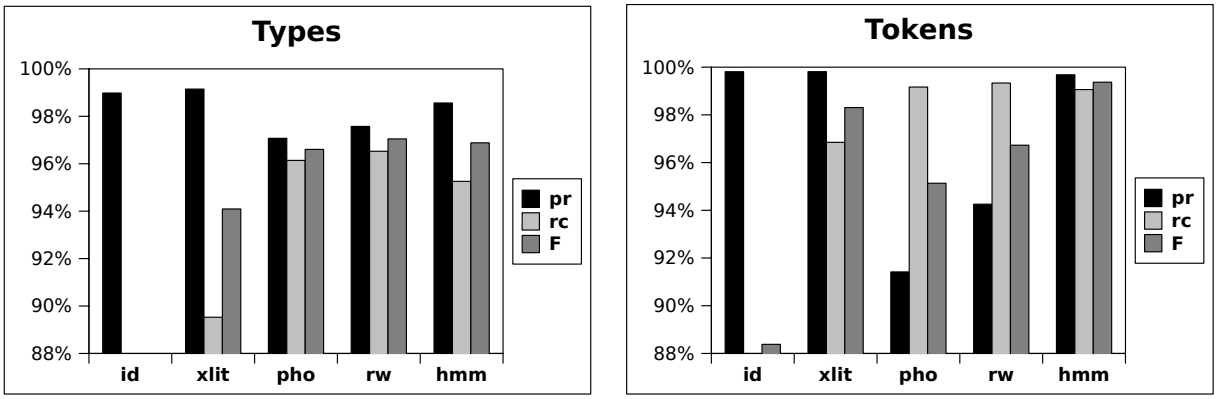

Figure 2: Evaluation data for various canonicalization techniques: visualization

1979) as a composite measure for both type- and token-wise evaluation modes:

$$\mathrm{F}(pr, rc) \quad = \quad \frac{2 \cdot pr \cdot rc}{pr + rc} \tag{23}$$

I follow Charniak et al. (1993) in using *relative error reduction rates* rather than absolute differences when comparing the performance of different canonicalizers. The general form for the (relative) error reduction in evaluation mode x provided by a method c_2 over method c_1 is: $\frac{x_{c_2} - x_{c_1}}{1 - x_{c_1}}$, assuming $0 \leq x_{c_1} \leq x_{c_2} \leq 1$. For example, given the data in Table 1, the error reduction in type-wise recall $x = \mathrm{rc}_{\mathrm{typ}}$ provided by $c_2 = \mathrm{rw}$ with respect to $c_1 = \mathrm{xlit}$ is $\frac{\mathrm{rc}_{\mathrm{typ,rw}} - \mathrm{rc}_{\mathrm{typ,xlit}}}{1 - \mathrm{rc}_{\mathrm{typ,xlit}}} = \frac{.965 - .895}{1 - .895} \approx 0.67 = 67\%$.

4.3 Results

Evaluation results for the canonicalization techniques described in sections 2 and 3 with respect to the test corpus are given in Table 1 and graphically depicted in Figure 2. Immediately apparent from the data is the typical precision–recall trade-off pattern

discussed above: conservative conflators such as string identity (id) and transliteration (xlit) have near-perfect precision ($\geq$ 99% both type- and token-wise), but relatively poor recall. On the other hand, ambitious conflators such as phonetic identity (pho) or the heuristic rewrite transducer (rw) reduce type-wise recall errors by over 66% and token-wise recall errors by over 75% with respect to transliteration, but these recall gains come at the expense of precision.

As hoped, the HMM disambiguator (hmm) presented in Section 3 does indeed recover a large degree of the precision lost by the ambitious type-wise conflators, achieving a reduction of over 41% of type-wise precision errors and of over 94% of token-wise precision errors with respect to the heuristic rewrite conflator. While some additional recall errors are made by the HMM, there are comparatively few of these, so that the type-wise harmonic average F falls by a mere 0.1% in absolute magnitude (3% relative error introduction) with respect to the highest-recall method (rw). Indeed, the token-wise composite measure F is substantially higher for the HMM disambiguator (99.4%, vs. 96.7% for the rewrite method), with an error reduction rate of over 64% compared to its closest competitor, deterministic transliteration (xlit).

The most surprising aspect of these results is the recall performance of the conservative transliterator xlit with $\mathrm{rc}_{\mathrm{tok}} = 96.8\%$, reducing token-wise recall errors by over 84% compared to the naïve string identity method. While such performance combined with the ease of implementation and computational efficiency of the transliteration method makes it very attractive at first glance, note that the test corpus was drawn from a comparatively recent text sample, whereas diachronically more heterogeneous corpora have been shown to be less amenable to such simple techniques (Gotscharek et al., 2009b; Jurish, 2010a).

5 Conclusion

I have identified a typical precision–recall trade-off pattern exhibited by several type-wise conflation techniques used to automatically discover extant canonical forms for historical German text. Conservative conflators such as string identity and transliteration return very precise results, but fail to associate many historical spelling variants with any appropriate contemporary cognate at all. More ambitious techniques such as conflation by phonetic form or heuristic rewrite transduction show a marked improvement in recall, but disappointingly poor precision. To address these problems, I proposed a method for disambiguating canonicalization hypotheses at the token level using sentential context to optimize the path probability of candidate canonical forms given the observed historical forms. The disambiguator uses a Hidden Markov Model whose lexical probabilities are dynamically re-computed for every input sentence based on the canonicalization hypotheses returned by a set of subordinated type-wise conflators, the entire canonicalization cascade remaining within the domain of weighted rational transductions.

The proposed disambiguation architecture was evaluated on an information retrieval task over a gold standard corpus of manually confirmed canonicalizations of historical

German text drawn from the *Deutsches Textarchiv*. Use of the token-wise disambiguator provided a relative precision error reduction of over 94% with respect to the best recall method, and a relative recall error reduction of over 71% with respect to the most precise method. Overall, the proposed disambiguation method performed best at the token level, achieving a token-wise harmonic precision-recall average $\mathrm{F} = 99.4\%$.

I am interested in verifying these results using larger and less homogeneous corpora than the test corpus used here, as well as extending the techniques described here to other languages and domains. In particular, I am interested in comparing the performance of the manually constructed rewrite transducer used here with a linguistically motivated language-independent conflator (Covington, 1996; Kondrak, 2000) on the one hand, and with conflators induced from a training sample by machine learning techniques (Ristad and Yianilos, 1998; Kempken et al., 2006; Ernst-Gerlach and Fuhr, 2006) on the other. Future work on the disambiguator itself should involve a systematic investigation of the effects of the various model parameters as well as more sophisticated smoothing techniques for handling previously unseen extant types and sparse training data.

Acknowledgements

The work described here was funded by a *Deutsche Forschungsgemeinschaft* (DFG) grant to the project *Deutsches Textarchiv*. Additionally, the author would like to thank Henriette Ast, Jörg Didakowski, Marko Drotschmann, Alexander Geyken, Susanne Haaf, Thomas Hanneforth, Wolfgang Seeker, Kay-Michael Würzner, and this paper's anonymous reviewers for their helpful feedback and comments.

References

Bahl, L. R., Jelinek, F., and Mercer, R. L. (1983). A Maximum Likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190.

Black, A. W. and Taylor, P. (1997). Festival speech synthesis system. Technical Report HCRC/TR-83, University of Edinburgh, Centre for Speech Technology Research.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of ANLP-2000*.

Brill, E. (1992). A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92*, pages 152–155.

Brill, E. and Moore, R. C. (2000). An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

Cafarella, M. and Cutting, D. (2004). Building Nutch: Open source search. *Queue*, 2(2):54–61.

Charniak, E., Hendrickson, C., Jacobson, N., and Perkowitz, M. (1993). Equations for part-of-speech tagging. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 784–789.

Church, K. W. and Gale, W. A. (1991). Probability scoring for spelling correction. *Statistics and Computing*, 1:93–103.

Covington, M. A. (1996). An algorithm to align words for historical comparison. *Computational Linguistics*, 22:481–496.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7:171–176.

DeRose, S. (1988). Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1):31–39.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

Ernst-Gerlach, A. and Fuhr, N. (2006). Generating search term variants for text collections with historic spellings. In Lalmas, M., MacFarlane, A., Rüger, S., Tombros, A., Tsikrika, T., and Yavlinsky, A., editors, *Advances in Information Retrieval*, volume 3936 of *Lecture Notes in Computer Science*, pages 49–60. Springer, Berlin / Heidelberg.

Gadd, T. N. (1988). 'Fisching fore werds': phonetic retrieval of written text in information systems. *Program*, 22(3):222–237.

Gadd, T. N. (1990). PHONIX: The algorithm. *Program*, 24(4):363–366.

Geyken, A. and Hanneforth, T. (2006). TAGH: A complete morphology for German based on weighted finite state automata. In *Proceedings FSMNLP 2005*, pages 55–66, Berlin. Springer.

Gotscharek, A., Neumann, A., Reffle, U., Ringlstetter, C., and Schulz, K. U. (2009a). Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, AND '09, pages 69–76, New York. ACM.

Gotscharek, A., Reffle, U., Ringlstetter, C., and Schulz, K. U. (2009b). On lexical resources for digitization of historical documents. In *Proceedings of the 9th ACM symposium on Document Engineering*, DocEng '09, pages 193–200, New York. ACM.

Hanneforth, T. and Würzner, K.-M. (2009). Statistical language models within the algebra of weighted rational languages. *Acta Cybernetica*, 19(2):313–356.

Jaynes, E. T. (1983). Brandeis lectures. In *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, pages 40–76. D. Reidel, Dordrecht.

Jelinek, F. and Mercer, R. L. (1980). Interpolated estimation of Markov source parameters from sparse data. In Gelsema, E. S. and Kanal, L. N., editors, *Pattern Recognition in Practice*, pages 381–397. North-Holland Publishing Company, Amsterdam.

Jelinek, F. and Mercer, R. L. (1985). Probability distribution estimation from sparse data. *IBM Technical Disclosure Bulletin*, 28:2591–2594.

Jurish, B. (2008). Finding canonical forms for historical German text. In Storrer, A., Geyken, A., Siebert, A., and Würzner, K.-M., editors, *Text Resources and Lexical Knowledge*, pages 27–37. Mouton de Gruyter, Berlin.

Jurish, B. (2010a). Comparing canonicalizations of historical German text. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology* (SIGMORPHON), pages 72–77.

Jurish, B. (2010b). Efficient online k-best lookup in weighted finite-state cascades. In Hanneforth, T. and Fanselow, G., editors, *Language and Logos: Studies in Theoretical and Computational Linguistics*, volume 72 of *Studia grammatica*, pages 313–327. Akademie Verlag, Berlin.

Keller, R. E. (1978). *The German Language*. Faber & Faber, London.

Kempken, S. (2005). *Bewertung von historischen und regionalen Schreibvarianten mit Hilfe von Abstandsmaßen*. Diploma thesis, Universität Duisburg-Essen.

Kempken, S., Luther, W., and Pilz, T. (2006). Comparison of distance measures for historical spelling variants. In Bramer, M., editor, *Artificial Intelligence in Theory and Practice*, pages 295–304. Springer, Boston.

Kernighan, M. D., Church, K. W., and Gale, W. A. (1990). A spelling correction program based on a noisy channel model. In *Proceedings COLING-1990*, volume 2, pages 205–210.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In *Proceedings NAACL*, pages 288–295.

Kondrak, G. (2002). *Algorithms for Language Reconstruction*. PhD thesis, University of Toronto.

Kukich, K. (1992). Techniques for automatically correcting words in texts. *ACM Computing Surveys*, 24(4):377–439.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(1966):707–710.

Lidstone, G. J. (1920). Note on the general case of the Bayes-Laplace formula for inductive or *a priori* probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Mays, E., Damerau, F. J., and Mercer, R. L. (1991). Context based spelling correction. *Information Processing & Management*, 27(5):517–522.

Möhler, G., Schweitzer, A., and Breitenbücher, M. (2001). *IMS German Festival manual, version 1.2*. Institute for Natural Language Processing, University of Stuttgart.

Mohri, M. (2002). Semiring frameworks and algorithms for shortest-distance problems. *Journal of Automata, Languages and Combinatorics*, 7(3):321–350.

Mohri, M. (2009). Weighted automata algorithms. In *Handbook of Weighted Automata*, Monographs in Theoretical Computer Science, pages 213–254. Springer, Berlin.

Pereira, F. C. N. and Riley, M. D. (1997). Speech recognition by composition of weighted finite automata. In Roche, E. and Schabes, Y., editors, *Finite-State Language Processing*, pages 431–453. MIT Press, Cambridge, MA.

Philips, L. (1990). Hanging on the metaphone. *Computer Language*, 7(12):39.

Philips, L. (2000). The double metaphone search algorithm. *C/C++ Users Journal*, June 2000.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Postel, H. J. (1969). Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. *IBM-Nachrichten*, 19:925–931.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286.

Rayson, P., Archer, D., and Smith, N. (2005). VARD versus Word: A comparsion of the UCREL variant detector and modern spell checkers on English historical corpora. In *Proceedings of the Corpus Linguistics 2005 conference*, Birmingham, UK.

Ristad, E. S. and Yianilos, P. N. (1998). Learning string edit distance. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 20(5):522–532.

Robertson, A. M. and Willett, P. (1993). A comparison of spelling-correction methods for the identification of word forms in historical text databases. *Literary and Linguistic Computing*, 8(3):143–152.

Russell, R. C. (1918). Soundex coding system. *United States Patent* 1,261,167.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Sokirko, A. (2003). A technical overview of DWDS/dialing concordance. Talk delivered at the meeting *Computational linguistics and intellectual technologies*, Protvino, Russia.

Unicode Consortium (2011). *The Unicode Standard*. The Unicode Consortium, Mountain View, CA.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA.

Verberne, S. (2002). *Context-sensitive spell checking based on word trigram probabilities*. Master thesis, University of Nijmegen.

Vidal, E., Thollard, F., de la Higuera, C., Casacuberta, F., and Carrasco, R. C. (2005). Probabilistic finite-state machines – Part II. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1026–1039.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, pages 260–269.

Zielinski, A., Simon, C., and Wittl, T. (2009). Morphisto: Service-oriented open source morphology for German. In Mahlow, C. and Piotrowski, M., editors, *State of the Art in Computational Morphology*, pages 64–75. Springer, Berlin.

Yvonne Zajontz, Marc Kuhn, Vanessa Kollmann

Markteffizienz durch Translation Memory Systeme? Intelligente Übersetzungstechnologien zur Reduktion von Transaktionskosten international agierender Unternehmen

Die Globalisierung der Märkte und die zunehmende Internationalisierung der Wirtschaft stellt Unternehmen vor enorme Herausforderungen. Neue Anforderungen in der wirtschaftlichen Information und Dokumentation haben immer größeren Einfluss auf den Außenhandelserfolg international agierender Unternehmen. Konsistente Terminologie- und Übersetzungsarbeiten können bisher nicht der Dynamik der internationalen Vernetzung der nationalen Volkswirtschaften folgen. Viele Unternehmen vernachlässigen diese erfolgskritischen Faktoren, obwohl (neue) Übersetzungstechnologien eine große Hilfe darstellen können. Welche Rolle spielen intelligente Übersetzungsspeichersysteme, sogenannte „Translation Memory Systeme“ (TMS), im Zuge der Globalisierung? Welche Anforderungen müssen diese Systeme aus Unternehmenssicht heute erfüllen, um Transaktionskosten senken zu können? Lassen sich aus der Sicht von Unternehmen weitere Optimierungspotentiale durch TMS-Produktinnovationen (wie etwa Open-Source-Lösungen) standardisieren? Dieser Beitrag versucht, die aufgeworfenen Fragen auf Basis empirischer Erhebungen zu beantworten.

New market requirements in the area of the economic information and documentation have a great impact on the foreign trade success of global acting companies. In this paper we briefly review new empirical research findings on the acceptance of Translation Memory Systems and the open source phenomenon in the field of translations and discuss the utility of Translation Memory Technology for the reduction of transaction costs. We offer some details regarding the requirements of Translation Memory Systems (TMS) and also offer some advice on conducting empirical studies on product innovation in the area of open TMS software.

1 Einleitung

Globalisierung und Internationalisierung gehören heute zu den am häufigsten verwendeten Stichworten wirtschafts- und gesellschaftspolitischer Diskurse. Während sich die Internationalisierung als Interimszustand in der Überleitung zur Globalisierung definiert und damit vor allem die geographische Expansion ökonomischer Aktivitäten über Ländergrenzen meint, handelt es sich bei der Globalisierung um eine weitaus fortgeschrittenere Entwicklungsstufe der Internationalisierung. Hinter dem Begriff verstecken

sich komplexe, hoch interaktive Vernetzungssysteme, sowohl auf politischer, soziokultureller als auch auf ökonomischer Ebene zwischen international agierenden Akteuren (vgl. HAAS & NEUMAIR 2005, S.3ff.). Die Globalisierung der Märkte und globale, arbeitsteilige Prozesse geben Anlass für zahlreiche wirtschaftswissenschaftliche Studien. Obwohl die immer „dichter“ werdende Welt sich in Lebenskultur und Verhaltensweisen zunehmend homogenisiert, spielen die Analyse der sprachlichen Diversifikation und die damit verbundenen Transaktionskosten durch sprachliche Standardisierungsmaßnahmen (Übersetzungen) im wissenschaftlichen Diskurs bisher eine eher untergeordnete Rolle. International agierende Unternehmen müssen den multilingualen Anforderungen in der wirtschaftlichen Information und Dokumentation gerecht werden. Wie gehen Industrie- und Übersetzungsunternehmen mit dieser Situation um? Die starke Zunahme des Fachdialoges infolge der Internationalisierung und Globalisierung über sprach- und geopolitische Grenzen hinaus hat die Übersetzungstätigkeit für Unternehmen zu einem wirtschaftlichen Faktor hinsichtlich Qualität und Rentabilität gemacht. Übersetzungsaufwendungen verursachen enorme Kosten. Ein Beispiel aus der Politik macht die Situation deutlich. So werden im EU-Parlament pro Jahr drei Millionen Seiten übersetzt. Die hohe Seitenanzahl lässt sich auf die 23 EU-Amtssprachen zurückführen. Eine zu übersetzende Seite kostet durchschnittlich 165 Euro. Hinzu kommen weitere Kosten für Simultanübersetzungen. Ein Sitzungstag im Europäischen Parlament kostet mehr als 90.000 Euro. 1.1 Milliarden Euro geben die europäischen Steuerzahler für Übersetzungstätigkeiten des EU-Parlaments jährlich aus, was circa einem Prozent des EU-Haushaltes entspricht (vgl. WELT ONLINE vom 28.6.2008). Automatisierte Übersetzungssysteme gewinnen aufgrund des gestiegenen Übersetzungsaufwandes zunehmend an Bedeutung (vgl. KÜDES 2002, S.12.). Dies gilt nicht nur für politische Institutionen, sondern auch für privatwirtschaftliche Unternehmen. Der vorliegende Artikel eruiert auf Basis empirischer Daten, welche Anforderungen Übersetzungstechnologien erfüllen müssen, um Transaktionskosten im Unternehmen zu minimieren. Gleichzeitig gibt der Beitrag Antwort auf die Frage, ob Unternehmen in übersetzungsbasierten Open-Source-Innovationen ein Optimierungspotential zur Senkung von Transaktionskosten erkennen.

2 Wettbewerbsvorteil durch international einheitliche Unternehmenssprache?

Um sich auf den internationalen Märkten durchsetzen zu können, spielen nicht nur Qualität und Preis von Produkten und Dienstleistungen tragende Rollen. Das Verhalten, die Kommunikation und das Erscheinungsbild eines Unternehmens müssen durchgängig und länderübergreifend verständlich sein, um sich auf dem Weltmarkt behaupten zu können. „Wer sofort erkennbar ist, wer ein klares Bild von sich abgibt, setzt sich am Markt durch“ (vgl. REINS 2006, S.9). Um seine Leistungen zu verkaufen, muss ein Unternehmen potentielle Kunden ansprechen. Die Ansprache setzt allerdings ein innersprachliches Verständnis zwischen Verkäufer und Käufer voraus. Qualitativ hochwertige Übersetzungsarbeit bildet die essentielle Basis für eine einheitliche internationale Unternehmenssprache. Fundierte Terminologiearbeit bestimmt die Qualität der Übersetzung, denn nur die Eindeutigkeit der Bezeichnung von Fachwörtern erleichtert den Dialog

innerhalb und zwischen Unternehmen. So kennzeichnen beispielsweise zwei unterschiedliche Bezeichnungen wie „Leichtmetallscheibenrad“ und „Alufelge“ denselben Gegenstand. Begriffliche Kohärenz ist jedoch notwendig, um sprachliche Missverständnisse zu verhindern. Sie fördert das Verständnis zwischen Interessensgruppen verschiedener Sprachräume und gewährleistet eine sprachliche Transparenz (vgl. KÜDES 2002, S. 9f.). Die Ausführungen haben gezeigt, dass die Terminologie die Einheitlichkeit von Begrifflichkeiten und Benennungen – für die interne und externe Unternehmenskommunikation den Erfolg des Unternehmens mitbestimmt. Konsequentes Terminologiemanagement in sämtlichen für das Unternehmen wichtigen Sprachen bildet die Basis für erfolgreiches Unternehmertum. Übersetzungsmanagement bildet daher einen Teil des unternehmensinternen Terminologiemanagements.

Gerade für ein Land wie Deutschland, das zum sechsten Mal hintereinander den Titel „Exportweltmeister“ von der Welthandelsorganisation WTO für das Jahr 2008 verliehen bekam, spielt effizient ablaufende Fachkommunikation in einer Vielzahl von Sprachen für den Wissensaustausch mit Handelspartnern und Kunden aus dem Ausland eine besondere Rolle (vgl. HUDETZ & FRIEDEWALD 2001, S.12). Wenn die Rezipienten nicht verstehen, was ihnen mitgeteilt wird, kann dies dem Unternehmen enormen Schaden zufügen. Abbruch von Handelsbeziehungen und Imageverluste sind nur einige der vielen negativen Auswirkungen von inkonsequentem Übersetzungsmanagement.

2.1 Übersetzen – bloß wie? Zur Organisation von Übersetzungen

In vielen (global agierenden) Unternehmen in Deutschland fehlt es bisher an einem konsequenten Übersetzungsmanagement, was die empirischen Ergebnisse der nachfolgenden Abschnitte belegen. Für international tätige deutsche Unternehmen sollten einheitliche englische, französische und spanische Übersetzungen zum Standardrepertoire der externen Kommunikation gehören. Übersetzungstätigkeiten wie zum Beispiel von Werbeauftritten, Websites, Dokumentationen oder Produktdeklarationen kann ein Unternehmen in unterschiedlicher Art und Weise organisieren:

1. Outsourcing von Übersetzungen: Übersetzungsaufgaben des Unternehmens werden an Drittunternehmen ausgelagert. Übersetzungsdienstleister bieten Translationstätigkeiten für Unternehmen an. Sie übernehmen Übersetzungsaufträge, indem sie innerhalb einer vorgegebenen Zeit einen Ausgangstext in das Äquivalent der Zielsprache übertragen.

2. In-house-Übersetzungsabteilung: Die zentrale Übersetzungsabteilung mit meist mehreren ausgebildeten Übersetzern übernimmt die Koordination von allen Translationen im Unternehmen und führt Übersetzungen für alle Abteilungen im Unternehmen durch.

3. Einzelübersetzer im Unternehmen: Ein ausgebildeter Einzelübersetzer steht für Übersetzungstätigkeiten im Unternehmen zur Verfügung. Bei Niedrigauslastung übernimmt der Übersetzer zusätzliche Aufgaben außerhalb seiner Kernkompetenz.

4. „unprofessionelle“ Übersetzungsarbeit im Unternehmen: Das Unternehmen überträgt Übersetzungsaufgaben an Mitarbeiter des Unternehmens (v.a. Mitarbeiter im Marketing), die die Zielsprache des zu übersetzenden Dokuments (mehr oder weniger gut) beherrschen.

5. Individuelle Übersetzungstätigkeiten von Mitarbeitern des Unternehmens: Jeder Mitarbeiter mit (einigermaßen) ausreichender Übersetzungskompetenz nimmt Übersetzungsaufgaben wahr.

Die Organisation der Übersetzungsarbeit durch nicht ausgebildete Übersetzer kann dem Unternehmen Imageschäden aufgrund schlechter Übersetzungsqualität auf dem fremdsprachigen Markt zufügen. Die ungleiche Verwendung von Terminologien kann zu Missverständnissen führen und einen Vertrauensverlust in das Unternehmen bewirken. Gleichzeitig führen inkonsistente Terminologien aufgrund einer großen Anzahl an Rückfragen und dem damit verbundenen Bearbeitungsaufwand zu einem erhöhten Kostenaufwand für das Unternehmen. Selbst bei global agierenden Unternehmen, für die Übersetzungen zur Routinearbeit zählen, mangelt es an konsequenter Terminologiearbeit und an einem effektiven Übersetzungsmanagement. Übersetzungen werden oftmals von Niederlassungen im Zielland oder dort ansässigen Übersetzungsbüros durchgeführt (vgl. HUDETZ & FRIEDEWALD 2001). Eine interne Zusammenführung der Übersetzungen findet jedoch häufig nicht statt.

In-house-Übersetzungsabteilungen sind zwar permanent disponibel und verfügen über unternehmensspezifisches Fachwissen, können aber bei Niedrigauslastung zu hohen Fixkosten führen. Die Auslagerung von Übersetzungsaufgaben gehört zu der gängigsten Organisation. Qualitativ hochwertige fachsprachliche und translatorische Endprodukte haben auf dem Übersetzungsmarkt jedoch ihren Preis. Insbesondere die Übersetzung von Fachtermini erfordert für Übersetzer langwierige Recherchearbeit, was sich im Zeitaufwand und damit auch in den Übersetzungskosten niederschlägt. Vor allem für kleine und mittlere Unternehmen, die meist nur über ein geringes Budget für Übersetzungsarbeit verfügen, bedeuten die neuen Anforderungen in der mehrsprachigen wirtschaftlichen Information und Dokumentation eine enorme Zusatzbelastung. Ein ausgearbeitetes, konsequentes Übersetzungsmanagement hilft, Transaktionskosten langfristig zu senken. Eine effiziente Möglichkeit zur Implementierung von Übersetzungsmanagement bieten so genannte Übersetzungsspeichersysteme (Translation Memory Systeme (TMS)), die terminologische Streuungen vermeiden sollen die Übersetzungsarbeit erleichtern.

2.2 Von der automatischen Übersetzung zu computergestützten „intelligenten“ Übersetzungssystemen

Galten vor einigen Jahren automatische Übersetzungsprogramme als Innovation auf dem Übersetzermarkt, so haben sich zwischenzeitlich „intelligente“ computergestützte Übersetzungssysteme zu unentbehrlichen Werkzeugen der täglichen Übersetzungsarbeit entwickelt (vgl. MASSION 2005, S.13). Während automatische Übersetzungssysteme

die Texte vollautomatisch, d.h. ohne die Hilfe von Übersetzern in die gewünschte Zielsprache übersetzen, handelt es sich bei computergestützten Übersetzungen (englisch „Computer-assisted translation (CAT)) um Humanübersetzungen, die durch Computerprogramme unterstützt werden. Automatische Übersetzungen verwenden maschinelle Übersetzungssysteme (MT-Systeme), die jedoch keine zufriedenstellenden Übersetzungsergebnisse liefern können (vgl. Abbildung 1). Vielmehr produzieren MT-Systeme (zum Beispiel Langenscheidt T1, L&H Power Translator, Logos, Reverso, Altavista, Heisoft) Rohübersetzungen, die eine Nachbearbeitung durch einen Übersetzer erfordern. Viele Sprachen enthalten zahlreiche übersetzerische Stolpersteine, wie z.B. die Variationsmöglichkeiten der Wortstellung, die sprachspezifischen Eigenheiten oder die Bedeutungsnuancen der Wörter, die bei maschineller Übersetzung zu weniger guten Translationsresultaten führen. Die Entwicklung von MT-Systemen ab den 50er Jahren verschlang viele Millionen Dollar. Der nachhaltige Erfolg blieb jedoch bis heute aufgrund der mangelhaften Übersetzungsergebnisse aus.

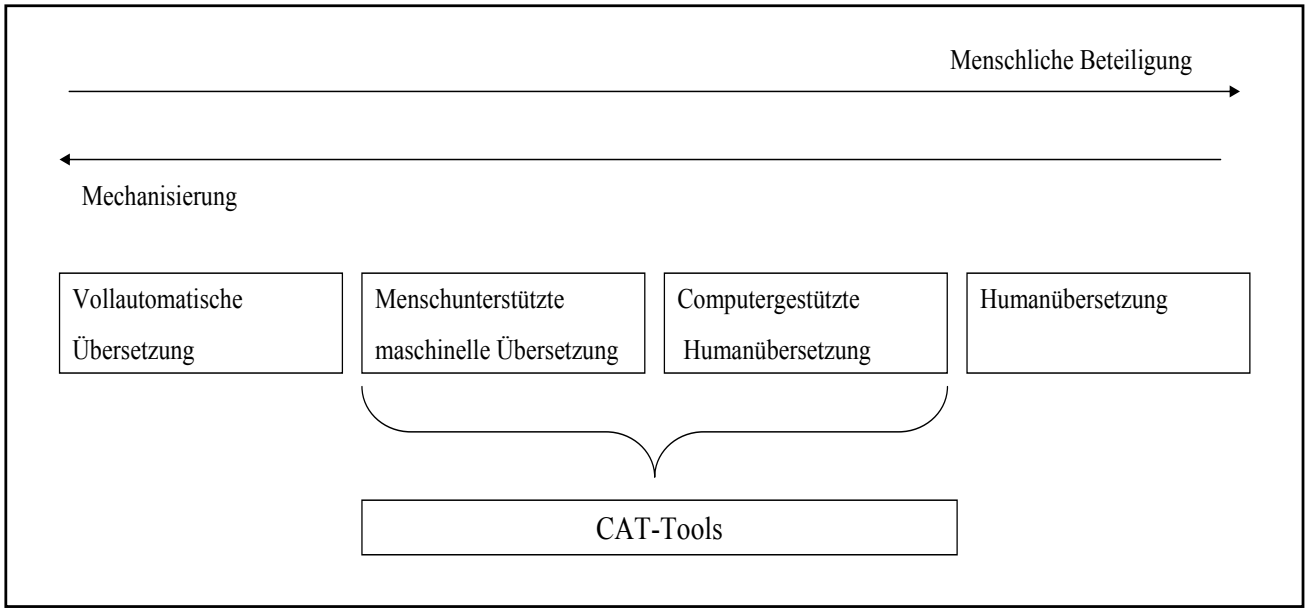

Abbildung 1: Einordnung von Übersetzungsmöglichkeiten (eigene Darstellung in Anlehnung an Hutchins & Somers 1992)

In den 80er Jahren besann man sich wieder auf die Kernkompetenzen der Translatoren und überließ die Übersetzungsarbeit den „menschlichen“ Übersetzern. Allerdings begann man damit, die Humanübersetzung durch computergestützte Systeme zu entlasten. Der Einsatz der CAT-Technologie führt dabei nicht nur zu Erleichterungen im Übersetzungsprozess, sondern zur möglichen Reduktion von Übersetzungskosten. Von den 80er Jahren bis heute hat die Funktionalität der CAT-Technologie stark zugenommen. Neue Komponenten wie die Verwaltung von Arbeitsabläufen und -prozessen, dezentrales Arbeiten und die Verwaltung von unterschiedlichen Dateiformaten ergänzen inzwischen die klassische CAT-Technologie (vgl. MASSION 2005, S.6f). Dennoch stoßen auch

computergestützte Werkzeuge an Grenzen. Sie erleichtern zwar die Übersetzungsarbeit, geben jedoch keine Qualitätsgarantie. Die Qualität wird von der geleisteten Recherchearbeit des Übersetzers oder Terminologen bestimmt, der das System aufbaut und pflegt.

CAT-Technologie hat sich vorwiegend im Bereich von Übersetzungsspeichern (englisch: Translation Memory System (TMS)) durchgesetzt. Bei TM-Systemen handelt es sich um Datenbanksysteme, die den zu übersetzenden Text zunächst in Übersetzungseinheiten (Segmente) zerlegen. Diese Segmente werden von einem Übersetzer übersetzt und im Translation Memory gespeichert. Dabei handelt es sich um eine Satzdatenbank, in der jeder bisher bearbeitete Satz zusammen mit seiner Übersetzung abgespeichert wird. Mit Hilfe eines Alignment-Tools können bereits übersetzte Segmente mit ihrer Übersetzung ins Translation Memory geladen werden. Bei neuen Übersetzungen werden die bereits existenten Vorübersetzungen zur Wiederverwendung vorgeschlagen. Das TMS markiert auch Segmente, die zwar nicht komplett mit bereits gespeicherten Übersetzungseinheiten übereinstimmen, aber gewisse Ähnlichkeiten aufweisen (so genannte Fuzzy-Matches) (vgl. MASSION 2005, S.15). Die Terminologiefunktion der TM-Systeme zeigt bei festgelegten Fachbegriffen die vorgegebene Übersetzung an. Damit gewährleistet das TM-System die Einheitlichkeit von Begrifflichkeiten und Benennungen. Der Übersetzer prüft im Anschluss die Verwendbarkeit der vorgeschlagenen Übersetzung. Er hat die Möglichkeit, das Segment für seine Übersetzung zu verwenden oder es gegebenenfalls zu überarbeiten. Je nach TM-System existieren weitere Zusatzfunktionen (zum Beispiel Qualitätssicherung, Statistik, Projektmanagement), die den Übersetzungsprozess zusätzlich erleichtern. In den vergangenen Jahren konnten sich zahlreiche Anbieter von TM-Systemen (u.a. Across, Déjà Vu, SDLX, Trados, Transit) auf dem Markt etablieren. Darüber hinaus machen es Übersetzungsspeichersysteme möglich, die terminologische und stilistische Konsistenz der Übersetzungen zu gewährleisten, was für die interne und externe Kommunikation eines Unternehmens von herausragender Bedeutung ist. CAT-Werkzeuge helfen ferner, neue Übersetzer in die Thematik einzuarbeiten und garantieren eine Rentabilität des investierten Kapitals, wenn komplexe technische Sachverhalte und hoch repetitive Dokumente zu übersetzen sind. Damit die Anschaffung von TM-Systemen in Unternehmen tatsächlich zu reduzierten Kosten beziehungsweise Transaktionskosten führt, müssen die Übersetzungstechnologien den Anforderungen des Unternehmens gerecht werden.

Vor diesem Hintergrund ergeben sich vier zentrale Fragen. Sind Übersetzungsspeichersysteme im Zuge der Globalisierung inzwischen – auch für Unternehmen – zu unverzichtbaren Instrumenten geworden? Welche Anforderungen müssen TM-Systeme aus Unternehmenssicht heute erfüllen, um Transaktionskosten senken zu können? Lassen sich aus der Sicht von Unternehmen noch weitere Optimierungspotentiale durch Produktinnovationen im TMS-Bereich (wie etwa Open-Source-Lösungen) standardisieren? Können durch Innovationen Transaktionskosten derart gesenkt werden, dass Unternehmen bereit wären, in neue Lösungsansätze zu investieren? Der Beitrag versucht auf der Basis empirischer Ergebnisse, Antworten auf diese Fragen zu finden.

3 Zur Methodik

Die folgenden Ausführungen basieren auf der Verwendung von quantitativen Forschungsmethoden, schließen jedoch qualitative Erhebungstechniken mit ein. Neben der Durchführung einer Sekundärdatenanalyse wurde eine onlinebasierte, standardisierte Befragung durchgeführt. Der Fragebogen richtete sich ausschließlich an Übersetzer oder Personen, die Übersetzungen in Unternehmen (aus Deutschland) koordinieren. Trotz der Zusammenarbeit mit dem Verband für Open Language Tools (FOLT) sowie dem Verband Deutscher Maschinen- und Anlagenbauer (VDMA), der zahlreiche Unternehmensadressen mit Ansprechpartnern zur Verfügung stellte, lag die Rücklaufquote nur bei 18,4%. Dennoch erreichte die Stichprobengröße insgesamt 247 Unternehmen. Der Beteiligungsschwerpunkt der Befragungsteilnehmer lag bei Industrieunternehmen (45,4%), Einzelübersetzern (15,8%) und Softwareunternehmen (14,2%). Die Befragung gewährleistet einen repräsentativen Querschnitt, der anhand eines Quotensamples generiert wurde. Sowohl kleine und mittelständische als auch große Firmen beteiligten sich an der Umfrage. Im Anschluss an die quantitative Erhebung wurden mit vier ausgewählten Experten aus der Stichprobe problemzentrierte Telefoninterviews geführt. Diese basieren auf einem halbstandardisierten, nach thematischen Bereichen gegliederten Leitfaden. Der strukturierte Leitfaden enthielt neben festgelegten Schlüsselfragen zu TM-Systemen ergänzende und vertiefende Eventualfragen. Die triangulatorische Verknüpfung der quantitativen und der qualitativen Daten führen zu den im Anschluss dargelegten Ergebnissen.

4 Translation Memory Systeme – ein effizientes Übersetzungsmanagement?

4.1 Gestiegene Relevanz von Übersetzungsarbeiten im Unternehmen?

Unternehmen erkennen im Zuge der Globalisierung eine zunehmende Relevanz von Übersetzungstätigkeiten innerhalb des eigenen Unternehmens. Über 80% der Befragten gehen von einer Zunahme firmeninterner Übersetzungsarbeiten in der Zukunft aus. Um sich am internationalen Markt behaupten zu können, müssen auch kleinere und mittlere Unternehmen die neuen multilingualen Anforderungen in der wirtschaftlichen Information und Dokumentation erfüllen. Trotz Erkenntnis der gestiegenen Relevanz unternehmensinterner Translationstätigkeiten spielt die Qualität der Übersetzungen bisher eher eine untergeordnete Rolle. Das gilt insbesondere, wenn man die Ausbildungshintergründe der Personen betrachtet, die Übersetzungstätigkeiten im Unternehmen durchführen. Nur 40% der Unternehmen beschäftigen ausgebildete Übersetzer mit Hochschulabschluss. In Industrieunternehmen fällt die Prozentzahl mit 26% ausgebildeten Übersetzern noch weitaus geringer aus. Technische Redakteure, Sekretärinnen, Marketingexperten oder sonstige Mitarbeiter mit Sprachkenntnissen übernehmen oftmals die Übersetzungsarbeit. Während Großunternehmen häufig über eigene zentrale Übersetzungsabteilungen verfügen, greifen kleinere und mittlere Unternehmen meist auf Mitarbeiter ohne Übersetzungsausbildung zurück. Eine zentrale Steuerung der Übersetzungstätigkeit fehlt

oftmals. Große organisatorische und strukturelle Schwächen in der Administration von Übersetzungstätigkeiten bestimmen das Bild in kleineren und mittleren Unternehmen. „Übersetzungsmanagement läuft bei uns in vielen Bereichen chaotisch ab. [...] Wir verfügen zwar über zwei ausgebildete Übersetzer, die üben aber mittlerweile andere Tätigkeiten im Unternehmen aus“, so die Mitarbeiterin eines mittelständischen Unternehmens, die Übersetzungen im Hause koordiniert. Qualitätssicherung in Form einer Prüfung von Terminologien (zum Beispiel Fachbegriffe, Produktbezeichnungen etc.) findet meist nicht statt. Vielmehr kommt es zu terminologischen Streuungen innerhalb des Unternehmens, da Terminologiearbeit häufig weder organisiert noch koordiniert abläuft, was die Qualität der Fachkommunikation mindert. Diese macht heute rund ein Fünftel der gesamten Informationen aus, die infolge der Globalisierung mit Hilfe der neuen Kommunikationstechnologien innerhalb der Informationsgesellschaft ausgetauscht wird (vgl. KÜDES 2002, S.9). Für große Übersetzungsprojekte engagieren über 75% der Unternehmen externe Übersetzungsdienstleister. Dabei sollte im Optimalfall die Beauftragung externer Übersetzungsbüros über eine zentrale Übersetzungskoordinationsstelle im Unternehmen laufen, um repetierende Textübersetzungen zu vermeiden. Viele der Befragten verfügen weder über Inhouse-Übersetzungsabteilungen noch über ein zentrales firmeninternes Übersetzungskoordinationsbüro. Vielmehr steuern mehrere (informell bestimmte) Personen unterschiedlicher Abteilungen Übersetzungen, was die hohe Anzahl an Übersetzungskoordinatoren erklärt. Die Steuerung von Übersetzungen übernehmen bei 20% der befragten Unternehmen mehr als fünf Personen. 12% davon gaben sogar an, dass mehr als elf Personen Übersetzungen im Unternehmen koordinieren. Terminologische Abweichungen, Doppelarbeit und die damit verbundenen „Extrakosten“ stellen häufig die Folge dar. Der Einsatz von Sprachtechnologien für die Archivierung oder Parallelspeicherung von Übersetzungen und für die Erstellung von Wortkonkordanzen kann Übersetzungstätigkeiten erleichtern und eine bessere „Verzinsung“ des investierten Kapitals gewährleisten. Übersetzungstechnologien bieten heute die Möglichkeit, Abläufe effizienter zu gestalten und durch kürzere Bearbeitungszeiträume für Übersetzungsaufträge Kosten zu reduzieren. Inwieweit hat sich diese Sichtweise tatsächlich in Unternehmen durchgesetzt? Das folgende Kapitel versucht, Antworten auf diese Frage zu finden.

4.2 Translation Memory Systeme – unverzichtbare Instrumente?

Obwohl 85% der Unternehmen Translation Memory Systeme als unverzichtbare Instrumente in der professionellen Übersetzungsarbeit ansehen, nutzen bisher nur 62% der befragten Institutionen Übersetzungsspeichersysteme. Für 22% spielt die Einführung von TM-Systemen derzeit keinerlei Rolle. Der Verzicht auf Übersetzungsspeichersystemen geht bei zahlreichen Befragten auf ein Informationsdefizit im Hinblick auf Sprachtechnologien zurück. 57% nannten als Hauptgrund, warum sie bisher kein TM-System nutzen, die Vergabe von Übersetzungsaufträgen an externe Dienstleister. Ihnen fehlt die Kenntnis, dass TM-Systeme nicht selbst übersetzen, sondern lediglich über intelligente Such- und Speichermechanismen verfügen, die die Übersetzungssituationen für

Übersetzer erleichtern sollen (SEEWALD-HEEG 2005, S.2). Würde ein Unternehmen ein TM-System besitzen, könnte es die bereits in einer Datenbank gespeicherten Terminologien und übersetzten Segmente an den Übersetzungsdienstleister transferieren. Langwierige Recherchen blieben dem Übersetzer erspart und die Genauigkeit, Konsistenz und Einhaltung sprachlicher und formaler Standards wären gewährleistet. Obwohl viele die hohen Anschaffungskosten eines Terminologieverwaltungssystems als Barriere sehen und damit den Nichterwerb erklären (vgl. SCHNEIDER 2007, S.66), machen nur 23% der Befragten die Kosten dafür verantwortlich. Reduzierte Übersetzungsarbeiten im eigenen Unternehmen sehen 40% als Hinderungsgrund, in TM-Systeme zu investieren. Ein Widerspruch zeigt sich jedoch bei der Betrachtung der Übersetzungsaufwendungen in Unternehmen, die nicht über ein TM-System verfügen. 60% der Personen, die in Unternehmen ohne Übersetzungsspeichersysteme arbeiten, verbringen mehr als 50% ihrer täglichen Arbeitszeit mit Übersetzungen. Immerhin planen 22% der Unternehmen ohne bisher vorhandene sprachtechnologische Lösungen die Anschaffung eines TM-Systems. Dennoch lehnen 51% der befragten Unternehmen ohne Übersetzungsspeicherwerkzeuge Investitionen in computergestützte translatorische Systeme weiterhin ab. Professionelle Übersetzungsarbeit erfordert ein professionelles Arbeitsumfeld und damit, wie der vorliegende Beitrag zeigt, ein leistungsfähiges Translation Memory System. Obwohl viele international agierende Unternehmen zahlreiche Anstrengungen bezüglich der Internationalisierung des Managements, der Erhöhung der internationalen Wertschöpfungsaktivitäten sowie der Führung und Organisation eines Netzwerkes ökonomischer Aktivitäten im internationalen Rahmen unternehmen, fehlt es oftmals an den Grundvoraussetzungen einer konsequenten Internationalisierungsstrategie – einem einheitlichen (länderübergreifendes) translatorischen Terminologiemanagement.

4.3 Anforderungen vs. Erfüllung – die Leistung von TM-Systemen

Für Unternehmen, die bereits mit Translation Memory Systemen arbeiten, bedeutet die Anschaffung von Übersetzungsspeichersystemen nicht die Lösung aller Übersetzungsmanagementprobleme. Vielmehr bedarf es einer konsequenten Steuerung und Organisation von Übersetzungsprozessen sowie einer Dokumentation und Archivierung von Übersetzungsarbeiten. Ein Translation Memory System ist nur so lange qualitativ hochwertig, wie dessen translatorische Inhalte auf einem hohen Niveau liegen. Ein effizientes System, dessen Daten sorgfältig gepflegt werden, ist deshalb äußert erfolgskritisch. Je mehr Übersetzer (mit unterschiedlichen Übersetzungsstilen) an dem System arbeiten, desto wichtiger wird eine konsequente Datenpflege. Die beste Technologie der Welt kann keine minderwertige Übersetzungsqualität kompensieren. So sieht es auch der Leiter der Übersetzungsabteilung eines mittelständischen Unternehmens: „Qualität ist so eine Sache. Wenn man zum Beispiel Englisch übersetzt, bekommt man immer einen anderen Übersetzer. Die haben dann natürlich unterschiedliche Übersetzungsstile. Da müssen wir uns dann halt darauf verlassen. Bei Sprachen, die nicht so häufig sind, hat man meist nur einen Übersetzer. Da ist die Qualitätswahrung dann eher gegeben.“ Ohne eine stringente Datenpflege wäre ein Übersetzungsspeichersystem für ein Unter-

nehmen weitgehend sinnlos. Dazu kommen Terminologien, die sich im Unternehmen immer wieder ändern oder doppelte Übersetzungen, die im System beseitigt werden müssen. Während die Datenpflege eine essentielle Grundvoraussetzung für einen effizienten Gebrauch von Translation Memory Systemen darstellt, muss darüber hinaus ein Übersetzungsspeichersystem für den Nutzer gewisse Anforderungen erfüllen, um ein effizientes Arbeiten gewährleisten zu können. Mittlerweile hat sich auf dem Markt der Translation Memory Systeme einiges bewegt. Der Anwender kann (theoretisch) je nach individueller Anforderung an das Übersetzungsspeichersystem zwischen verschiedenen TMS-Produkten wählen. TMS ist nicht gleich TMS. Derzeit existiert eine große Anzahl an Anbietern von Übersetzungsspeichersystemen, die unterschiedliche Schwerpunkte aufweisen. Sie unterscheiden sich einerseits in ihrer Technologie (Kompatibilität, Wörterbuchintegration, vernetzter Zugriff für Teams etc.) und ihrem Verwaltungsmanagement (integriertes Projektmanagementtool, Analysemöglichkeiten), andererseits in ihren translatorischen Produktionsverfahren (u.a. Automatisierung, effiziente Datenpflege, Qualitätskontrolle) sowie in ihren Ressourcen und Kosten (vgl. MASSION 2007, S.32). Trotz der Angebotsvielfalt fällt auf, dass drei Translation Memory Systeme unter den befragten Unternehmen dominieren. Während SCHNEIDER (2007, S.66) die Marktsituation als „Duopol“ beschreibt, zeigen die Ergebnisse der vorliegenden Studie vielmehr ein „Triopol“. Neben der klaren Dominanz von Trados (65%) heben sich weiter die TM-Systeme Transit (31%) und Accross (29%) von allen anderen Translation Memory Systemen ab. Da viele Nutzer mit verschiedenen Übersetzungsspeichersystemen arbeiten, besitzen inkompatible TM-Systeme eine geringere Absatzchance, obwohl sie aus technologischer Sicht oftmals den kompatiblen TM-Systemen vorzuziehen wären. Große Übersetzungsdienstleister besitzen häufig mehrere TM-Systeme, um den Anforderungen ihrer Kunden gerecht werden zu können. Meist erhält der Übersetzungsdienstleister trotz überdurchschnittlichem Fachwissen und Referenzen nur dann den Zuschlag, wenn er das passende TM-System zur Verfügung stellen kann. Kleinere Übersetzungsdienstleister oder Einzelübersetzer (welche in der Übersetzungsbranche stark dominieren) verfügen meist nur über ein spezielles TM-System. Dabei orientiert sich ihre Wahl des TM-Systems weniger an spezifischen Systemanforderungen, sondern weitaus mehr an den marktbeherrschenden Systemen, um sich dem Wettbewerb stellen zu können.

Das mag auch eine Erklärung für die (teilweis) großen Abweichungen zwischen den Anforderungen an ein TM-System (beziehungsweise der Wichtigkeit von Leistungen eines TMS) im Allgemeinen und der Erfüllung der Anforderungen durch das eigene System sein (vgl. Abbildung 2).

Die wichtigsten Leistungen, die ein TM-System aus der Sicht der befragten Unternehmen erbringen sollte, bestehen neben der großen Vielfalt an akzeptierten Dateiformaten in der parallelen Anzeige der ausgangs- und zielsprachlichen Segmente, dem sogenannte „Fuzzy Matching“ (Angabe von Ähnlichkeit zwischen dem im Translation Memory befindlichen ausgangsprachlichen und dem neu zu übersetzenden Segment in Prozent) und der Terminologiearbeit. Eine weitere wichtige Komponente bildet das Ex- und Importieren von Übersetzungsprojekten. Ein Import von Austauschformaten kann in manchen Systemen bis zu zwei Stunden in Anspruch nehmen (vgl. MASSION 2007, S.32). Wenn

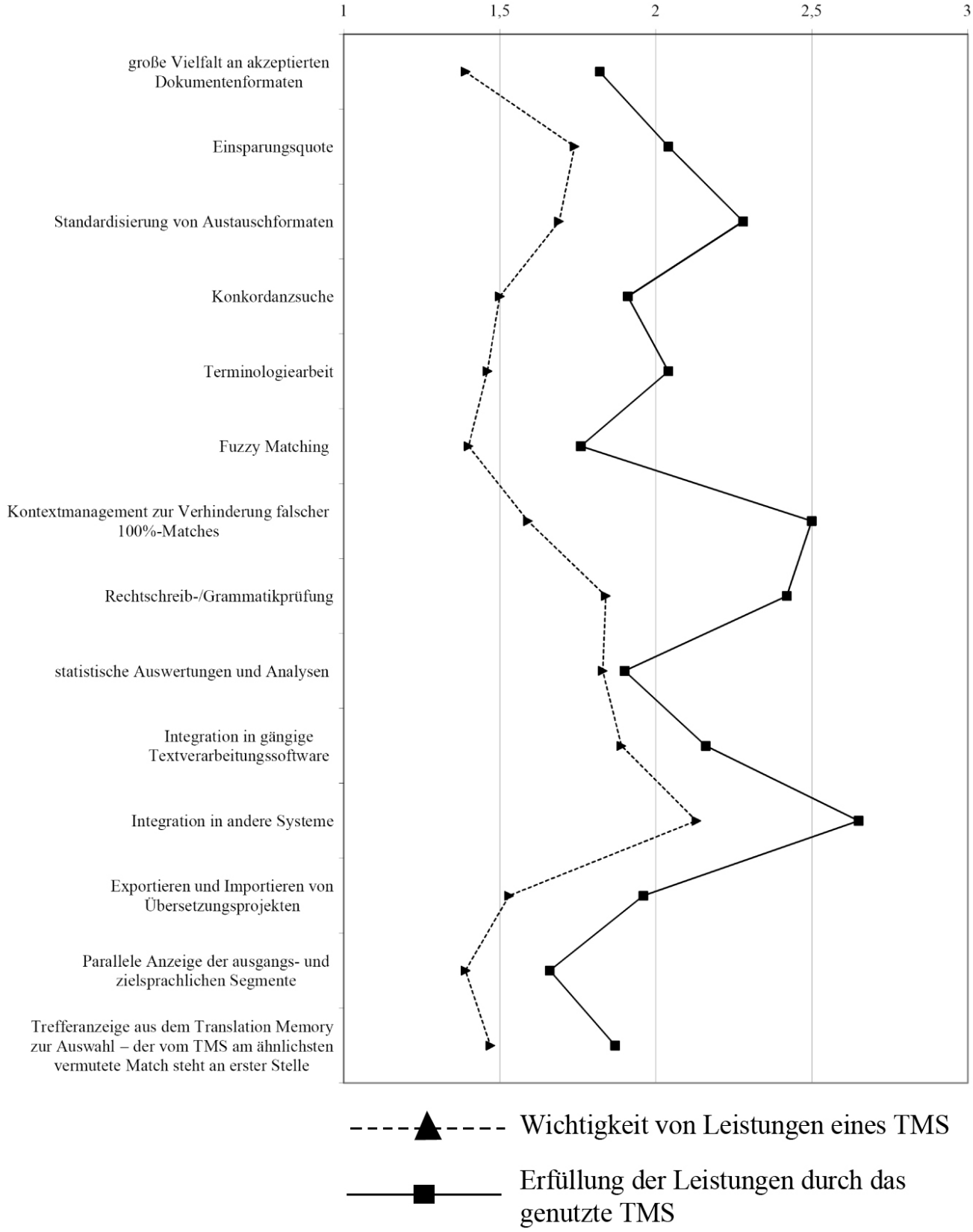

Abbildung 2: Wichtigkeit von Leistungen (1=sehr wichtig bis 4=unwichtig) eines TMS vs. Erfüllung der Leistungen durch genutztes TMS (1=sehr gut erfüllt bis 4=überhaupt nicht erfüllt) (eigene Erhebung)

in regelmäßigen Abständen Formate importiert werden, spielt die Importzeit für den Nutzer eine nicht zu vernachlässigende Rolle. Die Ergebnisse zeigen, dass zum großen Teil bedeutende Diskrepanzen zwischen den Anforderungen an TM-Systeme und deren Erfüllung durch die bestehenden Systeme bestehen. So konnte beim Kontextmanagement zur Verhinderung falscher 100%-Matches nur eine unzureichende Erfüllung durch das aktuell genutzte System festgestellt werden. Weitere Defizite bestehen darüber hinaus in den Bereichen der Terminologiearbeit und der Rechtschreib- und Grammatikprüfung. Nur in einigen wenigen Fällen erfüllen die aktuell von den Unternehmen genutzten TM-Systeme die Aufgaben bereits gut. Dazu gehören die statistischen Auswertungen und Analysen der Übersetzungstexte (zur Vorabschätzung des Übersetzungsaufwandes) und die Integration einer Textverarbeitungssoftware in das bestehende Übersetzungsspeichersystem. Bei diesen zwei Aspekten, die zur Qualitätssicherung beitragen sollen, ist zu beachten, dass ihre Wichtigkeit im Vergleich zu den anderen Punkten geringer ausfällt. In der Vergangenheit wurde die Qualitätssicherung von den Anbietern von Translation Memory Systemen eher stiefmütterlich behandelt, obwohl sie realistisch circa 15-20% der Produktionszeit eines Übersetzungsprojektes ausmacht. Standardfunktionen zur Qualitätskontrolle spielten in den jeweiligen TM-Systemen nur eine geringe Rolle, worin eine Erklärung für die Ergebnisse liegen könnte. Übersetzer griffen in der Vergangenheit vorwiegend auf entsprechende Qualitätssicherungstools zurück, die bislang nicht Bestandteil eines TM-Systems waren. Abbildung 3 zeigt deutlich, dass gerade die Qualitätssicherung unter den TM-Anwendern zu den wichtigsten Eigenschaften von Übersetzungsspeichersystemen zählt. Unter den Befragten, die bisher noch kein TM-System verwenden, zählen die Benutzerfreundlichkeit (62%) und die Rechtschreib- und Grammatikprüfung (50%) zu den wichtigsten Kriterien eines TM-Systems. Die Qualitätssicherung sehen nur 42% als ein sehr wichtiges Leistungsmerkmal eines TM-Systems an.

Obwohl viele TM-Systeme die Anforderungen der Nutzer nur unvollständig erfüllen, zeigen die Ergebnisse dennoch, dass Translation Memory Systeme die professionelle Übersetzungsarbeit effizient unterstützen können. Durch konsistente Übersetzungen heben sie die Übersetzungsqualität. Aufgrund der Zeitersparnis durch die Reduktion von Mehrfachübersetzungen können Übersetzer auch translatorische Dienstleistungen mit knappen Lieferterminen annehmen. Darüber hinaus stehen die in das System eingepflegten Daten zur Wiederverwendung für das nächste Übersetzungsprojekt zur Verfügung.

Die Kosten für die Anschaffung eines Translation Memory Systems bleiben dennoch der größte Kritikpunkt. Sie schlagen sich auf die Preise für die Übersetzungen nieder. Deshalb muss jedes Unternehmen die Wirtschaftlichkeit einer solchen Investition prüfen. So spielt nicht nur die Erstinvestition eine besondere Rolle bei der Kaufentscheidung, sondern es kommen für den Nutzer weitere Betriebskosten hinzu (u.a. Arbeitsaufwand mit dem Programm, Schulung von personellen Ressourcen, technische Betreuung, Lösen von Bugs) (vgl. MASSION 2007, S.34), die der Anwender einplanen muss. Während Freelancern häufig Sonder-konditionen bei der Anschaffung von TM-Systemen (häufig jedoch mit eingeschränkter Funktionalität) angeboten werden, müssen Unterneh-

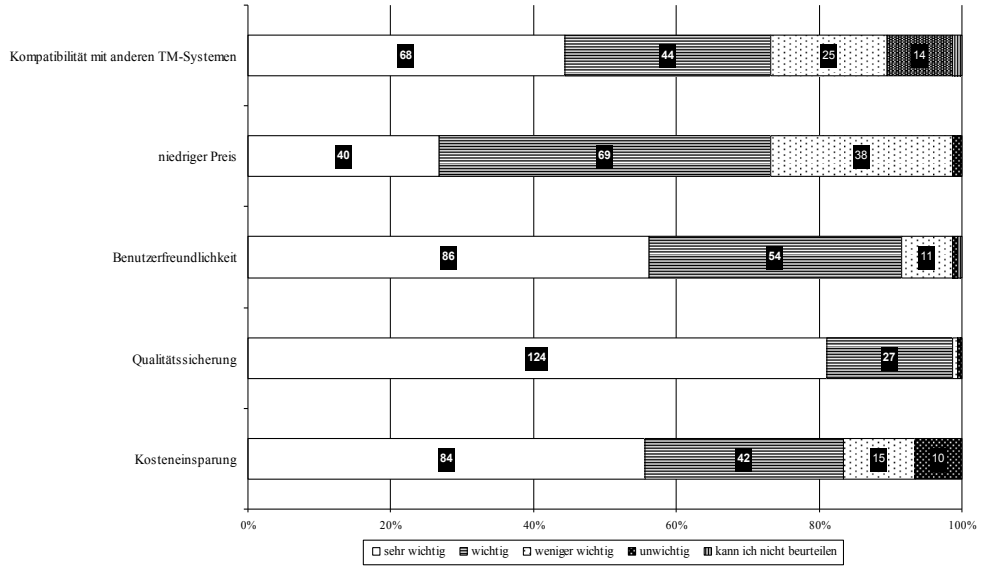

Abbildung 3: Wichtigkeit von Eigenschaften eines TM-Systems (unter TM-Nutzer) (eigene Erhebung)

men die Standardpreise einkalkulieren. Ob Kosten durch die Anschaffung eines TMS tatsächlich eingespart werden können (vgl. Abbildung 3), hängt von dem jeweiligen Übersetzungsaufwand im Unternehmen ab. Neue Konzepte und Entwicklungen könnten in den nächsten Jahren der schwerwiegenden Kostenkritik entgegenwirken. Könnten innovative Open-Source-Lösungen im TMS-Bereich weitere Optimierungspotentiale für den Anwender bringen?

4.4 Open-Source Übersetzungsspeicher-systeme – ein erfolgsversprechendes Modell?

Die historischen Wurzeln der Open-Source- und Free-Software-Bewegungen liegen in den 1980er Jahren (vgl. HELLER & NUSS 2004, S.386). Es gelang ihr, innerhalb weniger Jahre die Informations- und Kommunikationstechnologie stark zu beeinflussen und erhebliche Wirkungen auf dem Markt für Software zu erzielen. Unter Open-Source-Software wird eine Software verstanden, deren Quellcode offen gelegt wird und für den Anwender frei verfügbar ist. Open-Source-Software gewährt oftmals eine größere Sicherheit und Flexibilität als nicht frei zugängliche Programme bei hohem Entwicklungstempo und hoher Qualität. Geringe Kosten und schnelle, günstige Hilfe durch eine große Gemeinde von Open-Source-Entwicklern („Community“) bilden weitere Vorzüge einer frei verfügbaren Software. Aber lässt sich dieses Open-Source-Modell tatsächlich erfolgsversprechend auf Übersetzungsspeichersysteme anwenden? Nehmen Unternehmen eine derartige Produktinnovation wirklich an? Grundsätzlich stuften die Befragten Open TM-Systeme überwiegend als sehr vorteilhaft beziehungsweise vorteilhaft ein. Die größten Vorteile sehen Unternehmen in einem einfachen Handling durch die Nutzung eines einheitlichen Systems (42%), gefolgt von einem unkomplizierten Direktzugriff (41%) auf

das Programm und der Einsparung von Kosten durch die Reduktion der Systemvielfalt (41%). Die Verminderung beziehungsweise die Vermeidung von Folgekosten durch die Nutzung von Open TM-Systems beurteilen viele der Befragten (41%) als einen großen Vorteil. Um Open Source-Lösungen in ein verkaufbares Produkt zu verwandeln, bedarf es eines erfolgreichen Geschäftsmodells. Wesentlich in allen Open-Source-Debatten ist die freiwillige und zu einem großen Teil unentgeltliche Arbeit von Entwicklern und damit die Etablierung einer „Community“. Die Entwicklung, Markteinführung und Durchsetzung von „Open-Source-Lösungen“ sind somit in der Regel nur dann erfolgreich, wenn sich eine „kritische Masse“ an Marktteilnehmern an der Entstehung beteiligt, beziehungsweise diese fördert. Die befragten Unternehmen können sich insbesondere ein Engagement als Testuser der Software oder als Bereitsteller von Beratungsleistungen vorstellen. Darüber hinaus würden sich die befragten Unternehmen für die Bewerbung der Open TMS-Lösung im eigenen Wirkungsumfeld zur Verfügung stellen. Eine geringe Bereitschaft zeigen die Unternehmen bei der Bereitstellung von Serverkapazitäten und Entwicklerzeiten. Ein monetäres Engagement können sich viele der Befragten (70%) nicht vorstellen. Abbildung 4 zeigt das Interesse an Gegenleistungen der befragten Unternehmen für ein mögliches Engagement.

Die Ergebnisse machen deutlich, dass es bezüglich der Nutzung von lizenzfreien Systemen weiterhin auch große Vorbehalte gibt. Viele befragte Unternehmen sehen ein Defizit im Support und in der Qualität. Wer steht als kompetenter Ansprechpartner zur Verfügung? Wer kontrolliert das System? Kritisch stehen die Unternehmen auch dem Aufwand für Datenmigration gegenüber. Obwohl sich Open-Source-Lösungen in anderen Branchen vor allem durch schnellere Entwicklungen zu geringeren Kosten auszeichnen, sehen die Befragten gerade in der Entwicklungsgeschwindigkeit einen Nachteil von frei verfügbarer Translation Memory-Software. Mittlerweile existieren zahlreiche erfolgreiche Open-Source-Geschäftsmodelle. Um einen derartigen Entwicklungsstand zu erreichen, muss ein qualitativ hochwertiges, „offenes“ Produkt konstruiert und ein Nutzerkreis erschlossen werden. Die monetären Einnahmen generieren sich nicht wie bei vielen anderen Produkten aus dem Produkt selbst (im vorliegenden Fall die Open TMS-Software), sondern aus den dazugehörigen Dienstleistungen, die als Ware verkauft werden. Sie stellen ein proprietäres Gut des Open TMS-Software-Anbieters dar. Erst wenn diese Anforderungen erfüllt sind, können sowohl für die Anbieter als auch für die Anwender von Open Translation Memory Systemen Nutzen gezogen werden.

5 (Open) Translation Memory Systeme – ein System für die Zukunft? Ein abschließendes Fazit

Translation Memory Systeme werden bei steigendem Übersetzungsbedarf für ein Unternehmen immer wichtiger. Bei richtigem Einsatz können Translation Memory Systeme nicht nur eine verbesserte Qualität gewährleisten, sondern gleichzeitig Zeit sparen und Kosten reduzieren. Für intern-ational agierende Unternehmen entwickelten sich Übersetzungsspeichersysteme deshalb zu unverzichtbaren Instrumenten innerhalb der Organisation geworden. Die Befragungsergebnisse der Translation Memory System-Nutzer zeigen

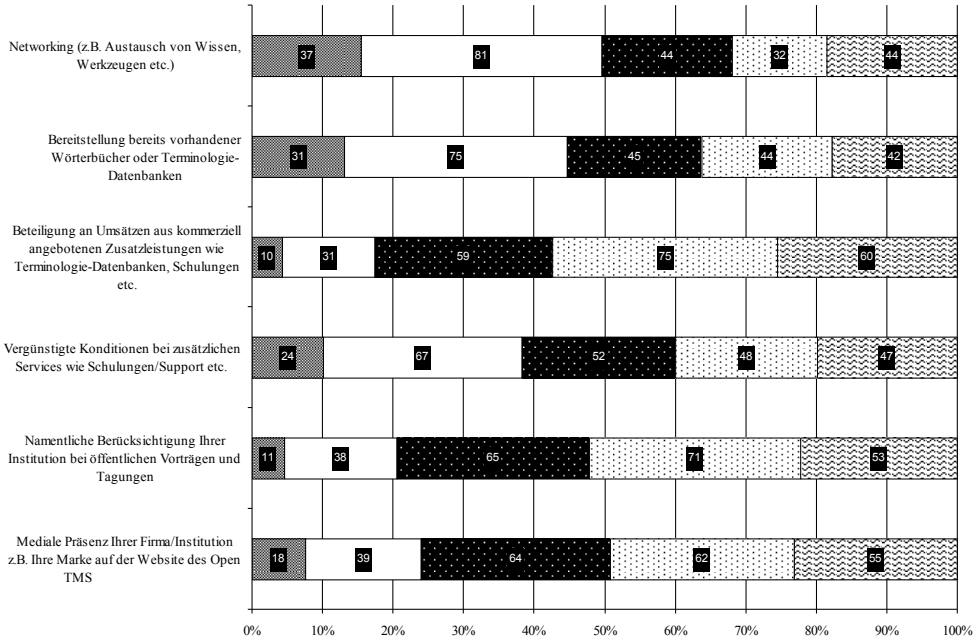

Abbildung 4: Interesse an Gegenleistung für Engagement bei Open TMS (eigene Erhebung)

allerdings einige Diskrepanzen zwischen den Anforderungen an TM-Systeme und deren Erfüllung durch die bestehenden Systeme. Die Interoperabilität zwischen den unterschiedlichen Translation Memory Systemen gilt als einer der größten Kritikpunkte unter den TMS-Anwendern. Drei Anbieter dominieren den Wettbewerb auf dem TMS-Markt. Aufgrund der Inkompatibilität der Systeme erhält meist der Übersetzungsdienstleister den Zuschlag, der das für das Unternehmen passende TM-System zur Verfügung stellen kann. Für Unternehmen, die bisher auf TMS-Lösungen verzichten, wäre die Anschaffung eines Systems denkbar, wenn die Anschaffungskosten sinken würden und das System darüber hinaus benutzerfreundlich konzipiert wäre. Durch innovative Open Source-Lösungen im TMS-Bereich könnten sich weitere Optimierungspotentiale ergeben. Die Interoperabilität der Werkzeuge oder die Verwendung offener und standardisierter Datenaustauschformate lässt sich durch Open TMS fördern. Unternehmen könnten dadurch unabhängig von bisherigen TMS-Herstellern werden. Kostenvorteile ebenso wie eine verbesserte Qualitätssicherung sollen die Produktinnovation gewährleisten. Dennoch zeigen die Ergebnisse, dass es bezüglich der Nutzung lizenzfreier Systeme Vorbehalte gibt. Existieren bei Produktin-novationen aber nicht grundsätzlich gewisse Vorbehalte? Doch gibt es die nicht immer bei Produktinnovationen? Die Bereitschaft Open TMS zu unterstützen, fällt bisher noch gering aus. Ein Engagement bei der Entwicklung von Open Source basierten TM-Lösungen können sich die Unternehmen überwiegend in nicht-monetären Bereichen vorstellen. Die Ergebnisse zeigen, dass es in den Unternehmen an Kenntnissen zu Übersetzungsspeichersystemen fehlt. Eine Sensibilisierung der Unternehmen für die Nutzung von TM-Systemen erscheint gerade im Zeitalter der Globalisierung als äußerst relevant. Vielleicht kann diese Informationsbeziehungsweise Sensibilisierungslücke von den Entwicklern der Open Source TMS-

Community genutzt werden, um auf ihre lizenzfreie Produktinnovation aufmerksam zu machen und um die bestehende Skepsis gegenüber TM-Systemen auszuräumen.

Literatur

H.-D. Haas, S.-M. N. (2006). Internationale Wirtschaft. Rahmenbedingungen, Akteure, räumliche Prozesse.

Heller, L. and Nuss, S. (2004). Open Source im Kapitalismus: Gute Idee - falsches System? In Lutterbeck, B. and Gehring, R. A., editors, *Open Source Jahrbuch 2004 - Zwischen Softwareentwicklung und Gesellschaftsmodell*, pages 385–405. Lehmanns Media, Berlin.

Hudetz, W. and Friedewald, M. (2002). Technische Produktdokumentation im Maschinen- und Anlagenbau. Eine Bestandsaufnahme.

Hutchins, W. J. and Somers, H. L. (1992). *An Introduction to Machine Translation*. Academic Press.

Küdes (Konferenz der Übersetzungsdienste europäischer Staaten) (2002). Empfehlungen für Terminologiearbeit.

Massion, F. (2005). Translation Memory Systeme im Vergleich.

Massion, F. (2007). Welcher Anbieter hat die besten Karten? TMS aus der Sicht eines Übersetzungsdienstleisters. In *Mitteilungen für Dolmetscher und Übersetzer*, pages 32–35.

Reins, A. (2006). Corporate Language: Wie Sprache Ã $\frac{1}{4}$ber Erfolg oder Misserfolg von Marken und Unternehmen entscheidet.

Schneider, M. (2007). Duopol als Ruhepol. Mehr Dynamik durch frei verfügbare TMS-Lösungen. In *Mitteilungen für Dolmetscher und Übersetzer*, page 66.

Seewald-Heeg, U. (2005). Der Einsatz von Translation Memory Systemen am Übersetzerarbeitsplatz. Aufbau, Funktionsweise und allgemeine Kaufkriterien. In *Mitteilungen für Dolmetscher und Übersetzer*, pages 8–36.

WeltOnline. Bundestag macht Bummelstreik gegen Brüssel.

Tim vor der Brück, Hermann Helbig

Meronymy Extraction Using An Automated Theorem Prover

In this paper we present a truly semantic-oriented approach for meronymy relation extraction. It directly operates, instead of syntactic trees or surface representations, on semantic networks (SNs). These SNs are derived from texts (in our case, the German Wikipedia) by a deep linguistic syntactico-semantic analysis. The extraction of meronym/holonym pairs is carried out by using, among other components, an automated theorem prover, whose work is based on a set of logical axioms. The corresponding algorithm is combined with a shallow approach enriched with semantic information. Through the employment of logical methods, the recall and precision of the semantic patterns pertinent to the extracted relations can be increased considerably.

1 Introduction

In most cases, objects are not elementary, rather composed of smaller objects, e.g., a car consists of wheels, windows, a gearshift, etc. Similarly, a group can be split up into its elements, e.g., a soccer team is composed of soccer players. Theses types of relationships are called meronymy. The whole or set is called the holonym while the corresponding part or element is called meronym.

Meronymy relations are required for a multitude of tasks in natural language processing, such as information retrieval or question answering. Let us consider a simple example. A user asks: "When was the last earthquake in Europe?". If the knowledge base contains the dates of recent earthquakes for all countries and also the information which countries are part (meronyms) of Europe, then this question can be answered.

To create large meronymy databases manually is very tedious and requires a lot of work. Thus, automatic approaches are preferable. A lot of approaches retrieve such relations by text mining. The first step is to develop a set of patterns. In the second step, these patterns are then applied to new texts, where they are used to recognize meronym/holonym pairs. Normally, these approaches only use surface or syntactical tree representations, i.e., constituency or dependency trees derived by a syntactical parser. They do not employ any semantic formalism and are therefore unable to incorporate background knowledge. Furthermore, they mostly extract meronyms between words and not word-readings. From this, it follows that results cannot be (directly) used in concept-based ontologies.

In this paper, a semantic approach is described which directly operates on SNs following the MultiNet (Helbig, 2006) formalism[1] in order to extract meronymy relations.

[1] MultiNet is the abbreviation of **Multi**layered Extended Semantic **Net**works.

This approach takes a knowledge base and a set of logical axioms into account. To this end, the entire content of the German Wikipedia with more than 20 million sentences is transformed into SNs using the WOCADI parser (Hartrumpf, 2002). The extraction patterns defined on a semantic level are mainly derived from the patterns given in (Girju et al., 2006). It is combined with a method based on shallow patterns enriched with semantic information if present.

In the next section, we review existing work on meronymy extraction. Section 3 describes the MultiNet formalism which is our representation for texts. An overview about the application of semantic patterns based on the MultiNet formalism is presented in Section 4. Section 5 describes how to incorporate a set of logical axioms and a knowledge base. The validation of extracted meronymy hypotheses is presented in Section 6. Section 7 illustrates how the correct meronymy subrelation can be selected. The architecture of our meronymy extraction system is given in Section 8. Evaluation results are specified in Section 9. Finally, the conclusion and an outlook of future work is given in Section 10.

2 Review of Existing Work

In this section, we give an overview on existing work on meronymy extraction. Quite popular are pattern-based approaches. Table 1 lists a collection of the patterns defined by Girju et al. (2006).

ID	Surface Pattern	Example
S_1	NP_{mero} is part of NP_{holo}	the **engine** is part of the **car**
S_2	NP_{holo}'s NP_{mero}	**girl**'s mouth
S_3	NP_{mero} of NP_{holo}	**eyes** of the **baby**
S_4	NP_{holo} verb:have NP_{mero}	The **table** has four **legs**.
S_5	NP_{holo} P NP_{mero}	A **bird** without **wings** cannot fly.
S_6	NP_{mero} P NP_{holo}	A **room** in the **house**.
S_7	$NP(N_{holo}\ N_{mero})$ (noun compound)	**door knob**
S_8	$NP(N_{mero}\ N_{holo})$ (noun compound)	**turkey pie**

Table 1: Some of the patterns suggested for the recognition of meronyms by Girju et al. (the lower indices mean: mero=meronym, holo=holonym)

These patterns are applied to arbitrary texts, and the instantiated variable NP_{mero} is extracted as meronym hypothesis of the assumed holonym NP_{holo}. Since hypotheses extracted by these patterns are not always correct, an additional validation component is required. Girju et al. employ a decision tree on annotated meronymy training data by making use of the WordNet hypernymy hierarchy.

Another pattern based approach is introduced by (Berland and Charniak, 1999). The validation of the extracted hypotheses is done by several statistical features taking the pattern by which a meronymy hypothesis was extracted into account as well as how likely the occurrence of the holonym hypothesis is if the given meronym hypothesis

shows up. In contrast to the approach of Girju et al., this method is not supervised and needs no annotated data.

Some of the patterns are very often applicable but the extracted hypotheses are rarely correct. An example of such a pattern is NP_{holo}'s NP_{mero}, as proposed by Girju et al. (2006). The ESPRESSO system introduced a bootstrapping approach geared towards handling this problem in particular. The reliability scores of relation hypotheses are used to derive reliability scores for the patterns that extracted such hypotheses and vice-versa (Pantel and Pennacchiotti, 2006).

An alternative approach to pattern matching is the use of support vector machines and tree kernel functions, which are employed to assign one of several semantic relations including meronymy to a given word or concept pair. A tree kernel function is a function for comparing trees, where the matrix of kernel values is symmetric and positive-semidefinite. Such approaches follow the assumption that a certain semantic relation (e.g., meronymy) is quite likely to hold if there exist a lot of sentences with similar tree structures (or similar paths in the dependency trees) in which this relation is known to hold (Culotta and Sorenson, 2004; Bunescu and Mooney, 2005; Zhao and Grishman, 2005; Reichartz et al., 2009).

Current approaches for meronymy extraction as described above are practically neither semantic-based nor do they take background knowledge into account. Let us consider two examples which demonstrate how background knowledge can improve the evaluation results.

The following pattern is given:

$$\text{MERO}(a1, a2) \leftarrow a1 \text{ is a member of } a2$$

This formula specifies that if a sentence contains the statement that *a1* is a member of *a2*, then *a1* is also a meronym (element) of *a2*. This pattern can be applied to the sentence *Mr. Peters is a member of AT&A.* to derive the meronymy relation[2] $\text{MERO}(Mr.Peters, AT\&T)$. Now consider the sentence: *Mr. Peters is the leader of AT&T.* Naturally, the pattern is not applicable to this sentence. However, if we use background knowledge, then the fact that *Mr. Peters is a member of AT&T* can eventually be inferred by the fact that *Mr. Peters is the leader of AT&T*, which makes the pattern applicable. Thus, a large knowledge base can reduce the number of required patterns considerably and therefore the amount of work for the pattern developer. In this example the knowledge base was employed to improve the recall but it is also possible to improve precision.

Consider a sentence matching the surface pattern *x is a mixture of y and z* (e.g., *Water is a mixture of hydrogen and oxygen.*). Two meronymy relations can be extracted from such a sentence: $\text{MERO}(y, x)$ and $\text{MERO}(z, x)$ (in the example $\text{MERO}(hydrogen, water)$ and $\text{MERO}(oxygen, water)$). Note that the word mixture can also be used in a more abstract sense, e.g., *His attitude is a mixture of enthusiasm and diligence.* In order to prevent in this case the extraction of assumed meronymy rela-

[2] This expression is not a valid MultiNet expression but stated rather informally.

tions like MERO(*enthusiasm*, *attitude*) and MERO(*diligence*, *attitude*), one has to require that x and/or y/z are known to be hyponyms of *substance*, which can be expressed by a logical constraint taking the transitivity axiom of hyponymy into account. This example is described in more detail in Section 5.

Finally, a knowledge base can be advantageous for a multitude of other tasks, e.g., the majority of the axioms presented here are also used for question answering.

3 MultiNet as a Fine-grained Semantic Network Formalism

As described in Section 2, axioms can be used to make patterns more generally usable or to support the specifications of logical constraints. Naturally, in order to use logical axioms, all sentences have to be converted into a logical representation. We have decided to use the MultiNet SN formalism since this is a logical representation with great expressiveness (even beyond first order predicate logic) and is excellently supported by several software tools. In contrast to networks such as GermaNet (Hamp and Feldweg, 1997) or WordNet (Fellbaum, 1998), MultiNet is designed to represent both semantic and lexical relations between lexems as well as the meaning of whole natural language expressions. An SN consists of nodes representing concepts (word readings) and arcs denoting relations between concepts or functions involving concepts. In total there are approximately 120 relations and functions defined in MultiNet, including the following:

- ARG1/2: Specification of relational arguments at the metalevel
- ATTCH: Attachment of an objects to another object
- ATTR/VAL: Attribute-value specification
- ELMT: Element relation
- HSIT: Relation specifying the constituents of a hyper-situation
- *ITMS: Function enumerating a set
- MERO: Meronymy relation, hyper-relation of ELMT, HSIT, ORIGM^{-1}, PARS, SUBM, and TEMP
- ORIGM: Relation of material origin
- PARS: Meronymy relation except ELMT, HSIT, ORIGM^{-1}, SUBM, and TEMP
- *PMOD: Modification of objects by associative or operational properties
- PRED: Predicative concept characterizing a plurality
- SUB: Relation of subordination for conceptual objects (hyponym/instance of)
- SUB0: Relation of general hyponymy, hyper-relation of SUB, SUBR and SUBS
- SUBM: Set inclusion (subset)
- SUBR: Relation of conceptual subordination for relations
- SUBS: Relation of conceptual subordination for situations
- TEMP: Relation specifying the temporal embedding of a relation

In MultiNet, concepts are specified by a word label and a pair of indices *.n.m* indicating the intended reading from a list of homographs or sememes of a polysemic word, respectively. These indices will henceforth be omitted from the text for the sake of brevity.

MultiNet is connected to the semantic lexicon HaGenLex (Hartrumpf et al., 2003). Each lexical entry of HaGenLex contains, aside from the typical morpho-syntactical information, one or more ontological sorts, a set of semantic features, and several layer features. The ontological sorts (currently more than 40) form a taxonomy. In contrast to other taxonomies, ontological sorts are not necessarily lexicalized, i.e., their names do not necessarily denote lexical entries.

The following list shows a small selection of ontological sorts:

- Object (o)
  - Concrete object (co): e.g., *milk*, *chair*
    - Discrete object (d): e.g., *chair*
    - Substance (s): e.g., *milk*, *honey*
  - Abstract object (ab): e.g., *thought*, *idea*
    - Abstract temporal object (ta): e.g., *month*
- Situation (si): e.g., *being warm*
- Quality (ql)
  - Property in the narrower sense (p): e.g., *tall*, *heavy*
  - Functional quality (fq): Such a quality obtains their full meaning only in connection with another entity.
    - Associative quality (aq), e.g., *chemical*, *philosophical*
    - Operational property (oq), e.g., *latter*, *third*

Semantic features denote semantic properties of objects; the values can be '+' (meaning applicable), '-' (not applicable) or 'underspecified'. A selection of semantic features is provided below:

- ANIMAL
- ANIMATE
- ARTIF (artificial)
- HUMAN
- SPATIAL
- THCONC (theoretical concept).

Sample characteristics of the nominal concept *bear*:

- Ontological sort: d (discrete object)
- Semantic features: ANIMAL +, ANIMATE +, ARTIF -, HUMAN -, SPATIAL +, THCONC -, ...

In this paper we only employ the layer feature *type of extensionality* (*etype*). Therefore only this feature is described. It classifies nodes on the pre-extensional knowledge representation level (see (Helbig, 2006) or (Lyons, 2002) for a distinction of intensional and (pre)extensional interpretation) and can assume one of the following values:

0: Representative of the extensional of an elementary concept, which is itself not a set, e.g., *house*, *Max* (person named Max)

1: Set of elements of type 0, e.g., *several children*, *three cars*, *team*, *brigade*

2: Set of elements of type 1, e.g., *three crews*, *many organizations*, *umbrella organization*

3: Set of elements of type 2, e.g., *three umbrella organizations*

This list can theoretically be continued to arbitrary type numbers but only types of extensionality until the value of three are realistic in practice.

The networks expressed in the MultiNet formalism are obtained from surface texts by means of the syntactico-semantic parser WOCADI (Hartrumpf, 2002), based on a word-class controlled functional analysis. Note that, although a meronymy relation can be represented in the MultiNet formalism, it is usually not contained in the SNs which are created by the parser unless such a relation is already comprised in the knowledge base.

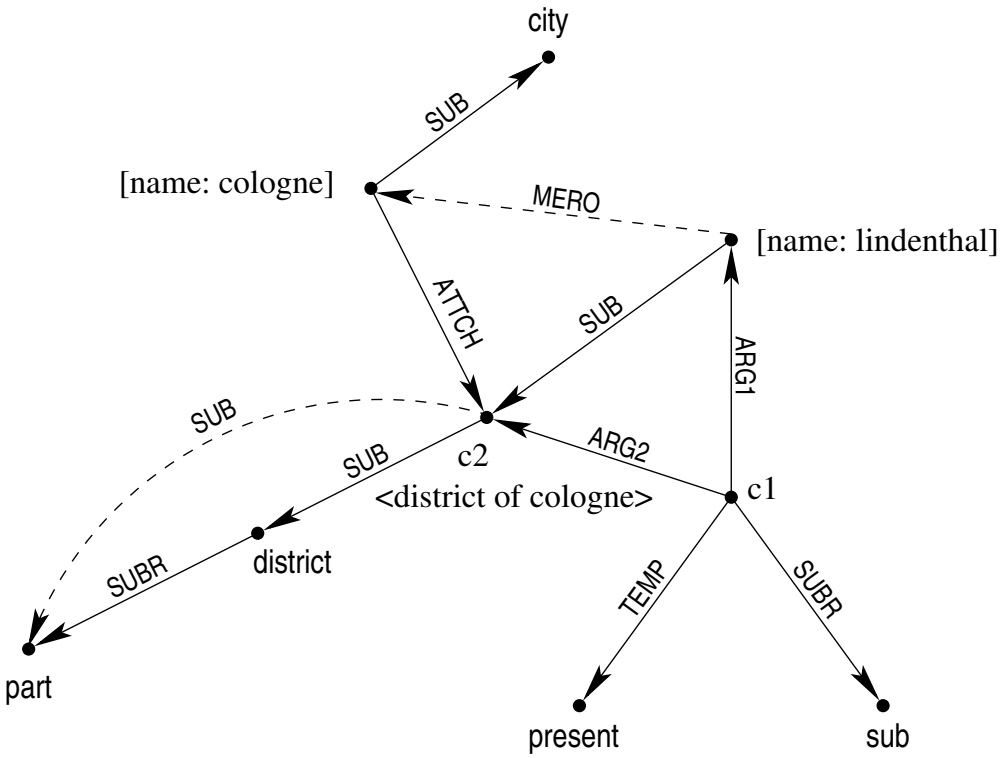

Figure 1: Application of the deep pattern D_4 to an SN representing the sentence: *Lindenthal is a district of Cologne.* Inferred edges are marked by dashed lines. SUBR$(c1, sub)$ indicates that the two arguments of $c1$ (c2 and c3) are connected by a SUB relation.

An example of an SN based on the MultiNet formalism is given[3] in Figure 1 and represents the sentence: *Lindenthal is a district of Cologne.* The specification of names (here Lindenthal and Cologne) is done by attribute/value constructs (MultiNet relations: ATTR and VAL). For better readability, an attribute value construct connected with a node c and subordinated to an attribute named n with value v, which is represented by the MultiNet expression ATTR$(c, d) \wedge$ SUB$(d, n) \wedge$ VAL(d, v), is written as $[n : v]$ in this figure. The parser recognized that a hyponymy relation is specified (SUBR(c1,sub)) and that the concept named *Lindenthal* is a hyponym (MultiNet relation: SUB) of the concept capsule associated to *district of Cologne.* The preposition *of* is realized by the relation ATTCH (attach) in the SN.

[3]Concept names are translated from German into English for better readability.

ID	Deep Pattern	Example
D_1	$\text{MERO}(a1, a2) \leftarrow \text{SUBS}(d, consist) \wedge \text{ARG1}(d, e) \wedge$ $\text{SUB}(e, a2) \wedge \text{ARG2}(d, f) \wedge$ $PAR_{*_{\text{ITMS}}}(f, g) \wedge \text{PRED}(g, a1)$	A car (a2) consists of wheels (a1),...
D_2	$\text{MERO}(a1, a2) \leftarrow \text{SUB}(c, a1) \wedge$ $\text{ATTCH}(d, c) \wedge \text{SUB}(d, a2)$	wheel(a1) of a car(a2)
D_3	$\text{MERO}(a1, a2) \leftarrow \text{ARG1}(e, d) \wedge \text{SUB}(d, a2) \wedge$ $\text{ARG2}(e, f) \wedge \text{SUB}(f, mixture) \wedge$ $\text{ATTCH}(g, f) \wedge PAR_{*_{\text{ITMS}}}(g, h) \wedge \text{SUB}(h, a1) \wedge$ $\text{SUBR}(e, sub) \wedge (\text{SUB}(a2, substance) \vee$ $\text{SUB}(a1, substance)) \wedge a2 \neq mixture$	Water (a2) is a mixture of hydrogen (a1) and ....
D_4	$N_1 := \text{ATTR}(a1, e) \wedge \text{SUB}(e, name) \wedge \text{VAL}(e, d)$ $N_2 := \text{ATTR}(a2, g) \wedge \text{SUB}(g, name) \wedge \text{VAL}(g, h)$ $\text{MERO}(a1, a2) \wedge N_1 \wedge N_2 \leftarrow$ $N_1 \wedge N_2 \wedge$ $\text{ARG1}(c, a1) \wedge \text{ARG2}(c, f) \wedge$ $\text{SUB}(f, part) \wedge \text{ATTCH}(a2, f) \wedge$ $\text{SUBR}(c, sub)$	Germany (a1) is part of Europe (a2).
D_5	$\text{MERO}(a1, a2) \leftarrow \text{ARG1}(d, e) \wedge$ $PAR_{*_{\text{ITMS}}}(e, f) \wedge \text{SUB}(f, a1) \wedge$ $\text{ARG2}(d, g) \wedge \text{PRED}(g, part) \wedge$ $\text{ATTCH}(h, g) \wedge \text{SUB}(g, a2) \wedge$ $\text{SUBR}(d, equ)$	Wheels, windows and a roof (a1) are part of car (a2).
D_6	$\text{MERO}(a1, a2) \leftarrow \text{SUB}(d, member) \wedge$ $\text{ATTCH}(e, d) \wedge \text{SUB}(e, a2) \wedge \text{ARG2}(f, d) \wedge$ $\text{ARG1}(f, g) \wedge \text{SUB}(g, d) \wedge \text{SUB}(g, a1)$	A goalkeeper (a1) is a member of a soccer-team (a2).

Table 2: A selection of deep meronymy patterns formulated by means of MultiNet relations ($PAR_f(x, y)$ denotes the fact that x is the result of function f where one of the arguments of f is y).

4 Application of Deep Meronymy Patterns

The meronymy extraction process is based on semantic patterns (see Table 2). Each pattern consists of a premise and a conclusion $\text{MERO}(a1, a2)$ (MERO is the MultiNet relation indicating meronymy) for generic concepts and $\text{MERO}(a1, a2) \wedge N_1 \wedge N_2$ for instances where N_1 and N_2 are attribute/value constructs for $a1$ and $a2$. The premise is given as an SN. Two of the nodes in this SN should be labeled with $a1$ and $a2$ in order for the pattern be applicable.

Shallow pattern	Matching expression
MERO(*a1*, *a2*) ← a1 ((word “of”)) ? (((cat (art)))) a2	wheel (a1) of a car (a2)
MERO(*a1*, *a2*) ← a2 ((word “with”)) a1	house (a2) with balcony (a1)
MERO(*a1*, *a2*) ← a2 ((word “without”)) a1	bird (a2) without wings (a1)
MERO(*a1*, *a2*) ← a1 ((word “is”)) ((word “the”)) ((word “main component”) ((word “of”)) a2	A CPU (a1) is the main component of computers (a2).
MERO(*a1*, *a2*) ← a1 * (((word “,”)) ? (((cat (art)))) a1) ((word “and”)) a1 ((word “are”)) ((word “parts”)) ((word “of”)) ? (((cat (art)))) a2	CPU (a1), . . .are parts of a computer (a2)
MERO(*a1*, *a2*) ← a1 * (((word “,”)) ? (((cat (art)))) a1) ((word “and”)) ? (((cat (art)))) a ((word “are”)) ((word “components”)) ((word “of”)) ? (((cat (art)))) a2)	CPU (a1) and display card (a1) are components of computers (a2)
MERO(*a1*, *a2*) ← *a2* ((word “consists”)) ((word “of”)) a1 ? (* (((word “,”)) a1) ((word “and”)) a1))	a computer (a2) consists of transistors (a1),. . .

Table 3: Selection of shallow patterns for meronymy extraction. The expressions occurring in the patterns are translated from German to English.

The matching of the pattern with an SN is done by an automated theorem prover for first order predicate calculus. In comparison to ordinary pattern matching, this has the advantage that logical axioms can be included into the pattern-matching process. Note that this paper only describes the use of a theorem prover for extracting and not for validating meronymy hypotheses, which is, for instance done by Suchanek et al. (2009); vor der Brück and Stenzhorn (2010). In total, there are 19 deep patterns which are mainly derived from the patterns introduced by (Girju et al., 2006).

To apply a pattern of the form[4] MERO(*a1*, *a2*) ← *premise*, where both *a1* and *a2* must show up in the premise, one has to find the bindings I for *a1* and *a2*, which are

[4]Patterns for instances can be applied similarly.

required to cause the following formula to become a tautology:

$$\begin{aligned} &\textsc{mero}(a1^I, a2^I) \leftarrow \\ &((\textsc{mero}(a1^I, a2^I) \leftarrow premise^I) \wedge SNX) \end{aligned} \tag{1}$$

in which $SNX = SN \wedge KB$ (KB=knowledge base). The knowledge base contains a set of axioms as well as a large number of lexical and semantic relations. If variable bindings are determined successfully, than the relation $\textsc{mero}(a1^I, a2^I)$ is extracted as meronymy hypothesis.

The proof is found by deriving a contradiction (i.e., the empty clause) from the negated expression (1), using resolution:

$$\begin{aligned} \bot \equiv &(((\textsc{mero}(a1^I, a2^I) \leftarrow premise^I) \wedge SNX) \wedge \\ &\neg\textsc{mero}(a1^I, a2^I) \Leftrightarrow \\ &(\text{Use distributive law and } A \leftarrow B \equiv \neg B \vee A) \\ \bot \equiv &\neg premise^I \wedge SNX \wedge \neg\textsc{mero}(a1^I, a2^I) \Leftarrow \\ \bot \equiv &\neg premise^I \wedge SNX \end{aligned} \tag{2}$$

Thus, since the trivial case that $\textsc{mero}(a1^I, a2^I)$ already appears in the knowledge base, premise or SN should be disregarded, it is required to derive a contradiction from $\neg premise^I \wedge SNX$. This is done by showing that the empty clause can be obtained applying logical resolution on $\neg premise^I \wedge SNX$. For this purpose, the MultiNet theorem prover is employed, which is also successfully used in question-answering tasks (Glöckner, 2007) and is optimized for this SN scenario. In our tests, the MultiNet theorem prover was ten times faster than the well-known general purpose theorem prover E-KRHyper (Baumgartner et al., 2007).

For easier processing, functions with a variable number of arguments are converted into a set of binary relations. For each such function $x_p = f(x_1, \ldots, x_l)$ we create l relations $PAR_f(x_p, x_1)$,...,$PAR_f(x_p, x_l)$ to represent the parent-child relationships between the result and the arguments and $(l(l-1))/2$ relations to represent the sequence of the arguments: $FOLL_f(x_i, x_j) \Leftrightarrow i < j$. Thus, an example expression $res = {}^*\textsc{itms}(x_1, x_2, x_3)$ (the MultiNet relation *ITMS combines several concepts in a conjunction) can be replaced by the following formula:

$$\begin{aligned} &\underset{^*\textsc{itms}}{PAR}(res, x_1) \wedge \underset{^*\textsc{itms}}{PAR}(res, x_2) \wedge \underset{^*\textsc{itms}}{PAR}(res, x_3) \wedge \\ &\underset{^*\textsc{itms}}{FOLL}(x_1, x_2) \wedge \underset{^*\textsc{itms}}{FOLL}(x_1, x_3) \wedge \underset{^*\textsc{itms}}{FOLL}(x_2, x_3) \end{aligned} \tag{3}$$

In addition to deep patterns, several shallow patterns are employed. Instead of a SN the premise consists of a regular expression involving feature value structures. The features are

- word: surface word form
- lemmas: possible lemmas

- categories: possible categories
- parse-lemma: lemma disambiguated by the Word Sense Disambiguation of the parser
- parse-reading: concept disambiguated by the Word Sense Disambiguation of the parser

These feature value structures are tried to be unified with the token information list provided by the parser. The most important employed shallow patterns are given in table 3. The applicability of these patterns could be further improved by adding additional optional tokens (like articles or adjectives). However, this make the patterns more difficult to read and also increases the extraction time. This is a drawback to deep patterns where such optional parameters are not needed.

5 Support via Logical Axioms

The use of an automated theorem prover together with the axiomatic apparatus of the MultiNet formalism has the advantage that the number of deep patterns can be considerably reduced compared with the number of shallow patterns. The axioms were mostly already developed for the task of question-answering and are reused for meronymy extraction.

ID	Axiom	# Hypotheses
A_1	$\text{SUB}(x,s) \leftarrow \text{SUB}(x,p) \wedge$ $*_{\text{PMOD}}(p,q,s) \wedge sort(q) = oq$	25 576
A_2	$\text{SUB}(x,z) \leftarrow \text{SUB0}(x,y) \wedge \text{SUB}(y,z)$	14 258
A_3	$PAR_{*_{\text{ITMS}}}(a,d) \wedge \text{PRED}(d,c) \leftarrow$ $\text{PRED}(a,c) \wedge \neg PAR_{*_{\text{ITMS}}}(e,a)$	2 117
A_4	$\text{SUBS}(x,z) \leftarrow \text{SUBS}(x,y) \wedge$ $\text{SUBS}(y,z)$	564
A_5	$\text{ATTCH}(a,e) \leftarrow \text{LOC}(e,l) \wedge$ $\{*_{\text{IN}}(l,a) \vee *_{\text{AT}}(l,a)\} \wedge$ $\{\text{SUBS}(e,s) \vee \text{PREDS}(e,s)\}$	194
A_6	$\text{PRED}(x,s) \leftarrow \text{PRED}(x,p) \wedge$ $*_{\text{PMOD}}(p,q,s) \wedge sort(q) = oq$	81
A_7	$\text{SUB}(a,s) \leftarrow \{\text{AGT}(e,a) \vee \text{EXP}(e,a) \vee$ $\text{MEXP}(e,a)\} \wedge \text{CTXT}(e,c) \wedge \text{SUB}(c,s)$	57

Table 4: Selected MultiNet axioms and the number of extracted hypotheses.

Table 4 presents the most successful axioms together with the number of hypotheses extractions for which a certain axiom was required. A_1 from Table 4 is required most often, where the function $*_{\text{PMOD}}$ is used to combine a conceptual object s with an operational property[5] (see Section 3), denoted by q in axiom A_1, which yields a more

[5]Operational properties having sort oq and associative properties are in contrast to properties in the narrower sense, which are treated by the MultiNet relation PROP (Helbig, 2006)).

special object p. This axiom is needed, for example, to deduce that *the first violinist* is a *violinist*.

Another example illustrating the usefulness of axioms is axiom A_2 (see Table 4). This axiom states the (generalized) transitivity of the hyponymy relation. Let us consider the SN representing the sentence *Lindenthal is a district of Cologne.* as displayed in Figure 1. Let us further assume the knowledge base contains the fact: SUBR(*district*, *part*). The concept represented by node *c2* is a hyponym of *district*. It can thus be inferred that *c2* is also a hyponym of *part*. Thus, the deep pattern D_4 (see Table 2) corresponding roughly to the surface pattern *a1 is a part of a2*, can also be applied to this SN. This means that the pattern D_4 becomes more generally usable by using axiom A_2.

A further frequently used axiom is

$$A_3 : \underset{*\text{ITMS}}{PAR}(a,d) \wedge \text{PRED}(d,c) \leftarrow \text{PRED}(a,c) \wedge \neg \underset{*\text{ITMS}}{PAR}(e,a) \tag{4}$$

This axiom is employed to add an *ITMS-relation to a PRED-edge, i.e., this axiom creates a coordination existing of exactly one element. In this way, a meronym inside and outside a coordination can be recognized by the same pattern.
Examples: **Car wheels** are part of a **car**. **Car wheels**, a **blinker** and a **gearshift** are part of a **car**. By applying axiom A_3 and pattern D_4, the meronymy relation MERO(*car_wheel*, *car*) can be extracted from both example sentences. The literal $\neg\, PAR_{*\text{ITMS}}(e,a)$ prevents the generated relation PRED(d,c) to be expanded by a further application of this axiom.

Axiom A_4 defines the transitivity of the SUBS-relation, and axiom A_5 states that an object located 'within' or 'at' another object is attached to this concept. A_6 is a variant of A_1; it only applies if pluralities (indicated by the MultiNet relation PRED) are considered. Finally A_7 is required if people act in certain roles (Example: *If someone does something as a father, then he is a father*).

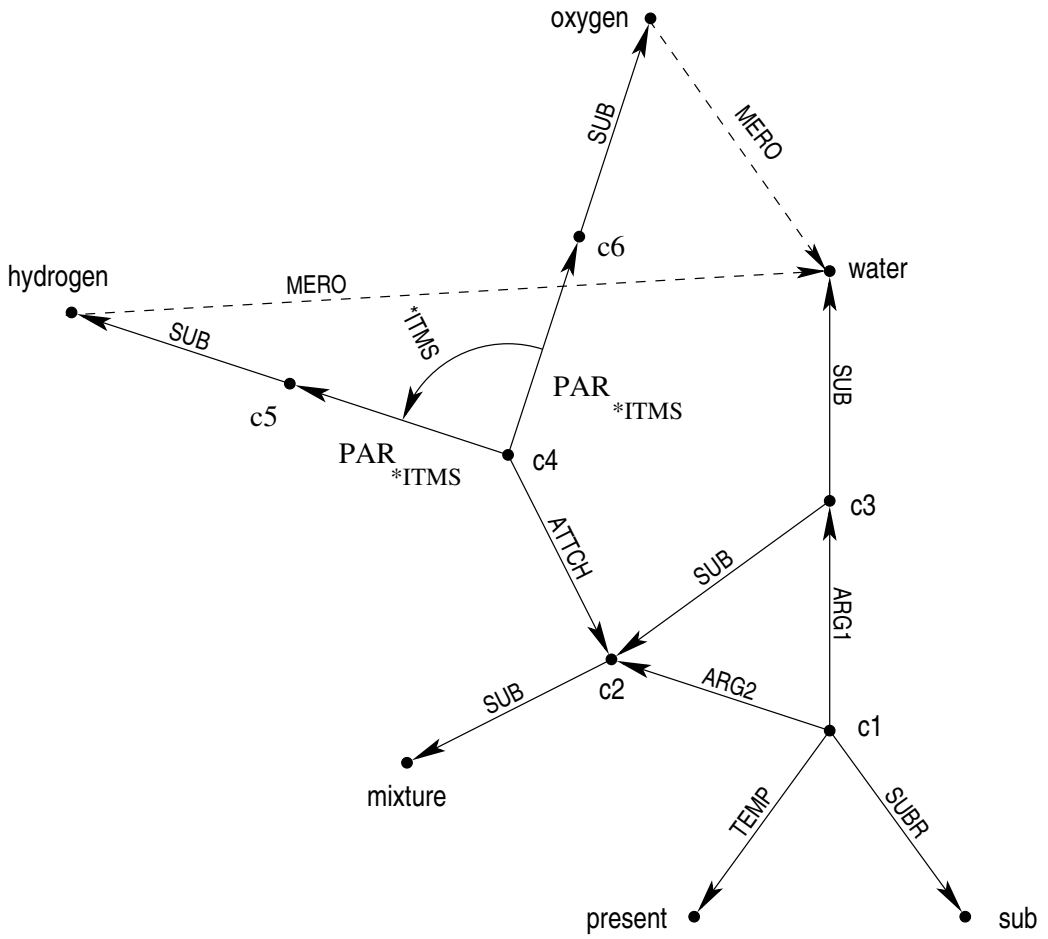

Figure 2: Application of the deep pattern D_3 to an SN representing the sentence: *Water is a mixture of hydrogen and oxygen.* The inferred edges are marked by dashed lines.

In the examples above, the use of axioms increases the recall. However, axioms can also help to increase the precision.

Consider the sentence *Water is a mixture of hydrogen and oxygen.* where the associated SN is given [6] in Figure 2. SUBR($c1, sub$) indicates that the two arguments of $c1$ (target concepts of ARG1 and ARG2) are connected by a SUB relation. The *ITMS function is used to combine the two components *hydrogen* and *oxygen* in a conjunction.

The two meronymy relations MERO($oxygen, water$) and MERO($hydrogen, water$) can be extracted from this sentence by applying pattern D_3 from Table 2 (roughly corresponding to the surface pattern given in Equation 5) to the associated SN.

$$\text{MERO}(x, z) \wedge \text{MERO}(y, z) \leftarrow z \text{ is a mixture of } x \text{ and } y \tag{5}$$

Note that the word mixture can also be used in a more abstract sense, e.g., *His attitude is a mixture of enthusiasm and diligence.* In order to prevent, in this

[6] Please note that c5 and c6 are generic nodes, and the arcs labeled with MERO should begin at c5 and c6, respectively. The latter is achieved in a post-processing step.

case, the extraction of assumed meronymy relations, like MERO(*enthusiasm*, *attitude*) and MERO(*diligence*, *attitude*), one has to require that at least one of y/z and x is known to be a *substance*, which is expressed by the disjunction SUB($a2$, *substance*) $\vee$ SUB($a1$, *substance*) in pattern D_3. A disjunction is used instead of the conjunction SUB($a2$, *substance*) $\wedge$ SUB($a1$, *substance*) since the lexical resources are limited and the hypernymy relations in the knowledge base are by no means complete. The pattern D_3 is applicable to the sentence *Water is a mixture of hydrogen and oxygen.* since water is a *substance*. The fact that water is a substance is derived by several applications of the axiom A_2 – *Transitivity of* SUB. The logical restriction $a_2 \neq$ *mixture* is required in order to prevent the extraction of MERO(*oxygen*, *mixture*) and MERO(*hydrogen*, *mixture*).

This pattern is not applicable in the aforementioned example, where *mixture* is used in an abstract sense, because neither *attitude*, *diligence* nor *enthusiasm* are hyponyms of *substance*.

6 Validation

Not all of the extracted meronymy hypotheses extracted by deep or shallow patterns are correct. Thus, a validation component is required which checks each hypothesis for correctness by several semantic and statistical features.

The knowledge validation carried out is done in two steps. In the first step we compare the ontological sorts of relational arguments and semantic features to each other and filter out hypotheses of non-allowed combinations. In the second step the remaining hypotheses are assigned a confidence score which estimates their probability of correctness. This two step mechanism was chosen for performance reasons. In this way the size of the database containing the meronymy hypotheses and the runtime of the confidence score computation can be greatly reduced.

Since the confidence score represents a probability, hypotheses with a score of at least 0.5 are considered to be correct. All other hypotheses are considered incorrect. This score can be useful for two reasons:

- Hypotheses with a score beyond a certain score threshold could automatically be added to the knowledge base.
- If the hypotheses are to be validated manually, the annotator can first check the hypotheses with the high scores. In this way, he can add much more meronyms to the knowledge base in a given time interval than if he chooses the hypotheses randomly.

6.1 Filtering

Only certain combinations of ontological sorts and semantic features are permitted. For instance, a concept denoting a human being (semantic feature: human +) can only be meronym to another concept that denotes a set (recognizable in MultiNet by the type of extensionality). The type of extensionality is zero for an individual concept (which is not a set) and i for a set of elements of type $i-1$ for $i > 0$ (see Section 3). The

ontological sorts of the meronym and holonym hypothesis must usually be identical (exception: sort s (substance) for meronym and d (discrete) for holonym or vice-versa are allowed).

The permitted combinations of ontological sorts are specified manually, while the regularities concerning features, which are a lot more complicated, are automatically learned by a tree augmented naïve Bayes algorithm (TAN) (Friedmann et al., 1997). The training is done separately for different combinations of the type of extensionality. The training data consists of a set of annotated meronymy relation hypotheses (meronym/no meronym) and lists of semantic features where the features are sought and found in the lexicon. Only relational candidates for whom semantic features and ontological sorts can be shown to be compatible are stored in the knowledge base.

6.2 Scoring

All hypotheses in the knowledge base are assigned a confidence score. This is done by means of a support vector machine (SVM) applied on several feature values[7] and annotations. The SVM (here LIBSVM, Chang and Lin (2001)) determines the class (meronymy or non-meronymy) and a probability estimate for each candidate pair and is trained on a set of annotated meronymy hypotheses. The annotation is either one (1) for meronymy or zero for non-meronymy. If the classification is 'meronymy', the score is defined by this probability estimate, otherwise as one minus this value. The employed features are explained below:

Use of a Taxonomy: This feature exploits a collection of known meronyms, hyponyms, synonyms, as well as a list of known non-meronym pairs. For that we used the lexical and semantic relations contained in HaGenLex. These relations are in part handcrafted, and in part derived from Wiktionary. Additionally, GermaNet relations, where the synsets are mapped to HaGenLex concept ids, are employed. This method works as follows: Consider a given pair of concepts $(a1, a2)$ being meronymically related to each other. Determine the Cartesian product $S(a1) \times S(a2)$ of all hypernyms of both normalized components (including the components itself). Increase a counter *pos* for all possible pairs of concepts in the resulting Cartesian product $S(a1) \times S(a2)$, i.e., set

$$\begin{aligned} pos(x,y) :=& pos(x,y) + 1 \\ & \forall x \in S(a1), y \in S(a2) \end{aligned} \tag{6}$$

where

$$S(x) = \{syno\_normalize(x)\} \cup \{syno\_normalize(z) | \textsc{sub}(x,z)\} \tag{7}$$

[7] not to be confused with semantic features

where $syno\_normalize : Concepts \rightarrow Concepts$ is a function which maps a concept to the smallest element (according to some total ordering) of its synset:

$$syno\_normalize(c) = d :\Leftrightarrow \text{SYNO}(d, c) \wedge (\forall e : \text{SYNO}(e, d) \Rightarrow e \geq d) \tag{8}$$

Example: Assume $synset(\text{car})=\{\text{auto,car}\}$. Then $syno\_normalize(car) = auto$ if lexicographic ordering is used. By employing a synonymy normalization, the set of regarded concepts can be reduced which leads to a smaller memory consumption.

Analogously to *pos*, determine $neg(x, y)$ for all non-meronym pairs. For a new pair (x', y') opt for meronymy, iff

$$\max_{a \in S(x'), b \in S(y')} \left\{ \frac{pos(a,b)}{pos(a,b)+neg(a,b)} \right\} > \max_{c \in S(x'), d \in S(y')} \left\{ \frac{neg(c,d)}{pos(c,d)+neg(c,d)} \right\} \tag{9}$$

That means that the decision is in favor of meronymy iff there is a pair in $S(x') \times S(y')$ for which the indication for meronymy is stronger than the indication against meronymy for any other pair in $S(x') \times S(y')$. This decision procedure basically follows the approach presented by Costello (2007).

In addition to the approach of Costello, we employed the taxonomy to cancel out all meronymy hypotheses that appear in or can be derived from this taxonomy since a meronym cannot be a hyponym or a hypernym simultaneously. If one or both of the considered concepts are associated to compound words, then we also check for known hyponymy relations between all combinations of meronymy concept+base concept and holonymy+base concept (base concept: concept corresponding to the correct reading of the base word) candidates.

Correctness Rate: The feature *Correctness Rate* takes into account that the recognized holonym alone is already a strong indication for the correctness or incorrectness of the investigated hypothesis. The same holds for the assumed meronym as well. For instance, meronymy candidate pairs with the assumed holonym *computer*, *university*, or *building* were mostly correct. In contrast, meronymy candidate pairs with an assumed holonym *future* or *reason* were usually incorrect.

Thus, this indicator determines how often a concept pair is actually correct if a certain concept shows up in the first (meronym) or second (holonym) position. More formally, we are interested in determining the following probability:

$$p = P(me = t | arg1_r = a1 \wedge arg2_r = a2) \tag{10}$$

where

- $arg1_r$ denotes the first concept (the assumed meronym) in a given relation r.
- $arg2_r$ denotes the second concept (the assumed holonym) in a given relation r.
- $me(ronym) = t(rue)$ denotes the fact that a meronymy relation holds.

Applying Bayes' theorem to Equation 10 leads to the Equation:

$$p = P(me = t) \cdot \frac{P(arg1_r = a1 \wedge arg2_r = a2 | me = t)}{P(arg1_r = a1 \wedge arg2_r = a2)} \tag{11}$$

For better generalization, we assume that the events $arg1_r = a1$ and $arg2_r = a2$ as well as $(arg1_r = a1 | me = t)$ and $(arg2_r = a2 | me = t)$ are independent. Using these assumptions, Equation 11 can be rewritten:

$$\begin{aligned} p \approx p' &= P(me = t) \cdot \frac{P(arg1_r = a1 | me = t)}{P(arg1_r = a1)} \cdot \frac{P(arg2_r = a2 | me = t)}{P(arg2_r = a2)} \\ &= \frac{P(arg1_r = a1 \wedge me = t)}{P(arg1_r = a1)} \cdot \frac{P(arg2_r = a2 \wedge me = t)}{P(me = t) \cdot P(arg2_r = a2)} \\ p' &= \frac{1}{P(me = t)} \cdot P(me = t | arg1_r = a1) \cdot P(me = t | arg2_r = a2) \end{aligned} \tag{12}$$

If *a1* only rarely occurs in meronym position in assumed meronymy relations, we approximate p by $P(me = t | arg2_r = a2)$, analogously for rarely occurring concepts in the holonym position. As usual, the probabilities are estimated by relative frequencies relying on a human annotation. Example: Let us consider that a hypothesis pair (c_1, c_2) is given and concept c_1 occurs in 90 annotated meronymy hypotheses as an assumed meronym, 45 of them are known to be correct. The concept c_2 occurs in 100 hypotheses as holonym candidate and 30 of them are known to be correct. Let the probability that a meronym hypothesis is correct be 0.2. Then the total score is given as: $(1/0.2) \cdot 0.5 \cdot 0.3 = 0.75$.

Graph Kernel: The use of kernels (see Section 2) is quite popular for semantic relation extraction. Since relation extraction algorithms are mainly based on syntactic or surface structures, tree or string kernels are usually applied. In our scenario, the kernel is only applied for validation of hypotheses where the extraction is done by the automated theorem prover. Hence, a hybrid approach is taken. In addition, our method is based on SNs, which are graphs and not trees. Thus, instead of the usual tree kernel, a graph kernel (vor der Brück and Helbig, 2010) based on common walks, as proposed by (Gärtner et al., 2003) is applied.

Applied Pattern: There are major differences regarding the precision values of the extracted meronymy hypotheses depending on the applied pattern (Berland and Charniak, 1999). Pattern D_1 (see Table 2) is actually quite reliable where D_2 generates a lot of incorrect hypotheses. Thus, we provide a feature for each pattern which takes the value of one if a meronymy hypothesis was extracted by this pattern, to zero otherwise.

Mutual Information: Relation hypotheses extracted several times are often more reliable than hypotheses that could only be found once. Thus, we introduce a feature measuring the point-wise mutual information (in contrast to the conditional probability in Berland and Charniak (1999)) between the meronym and holonym candidate multiplied by the discounting factor suggested in (Pantel and Ravichandran, 2004).

Deep/Shallow: This feature checks whether a hypothesis was extracted by both shallow and deep patterns (1) or only by one of them (0).

Concrete/Abstract: This feature follows the assumption that most meronymically related concepts are concrete. It is the product of the concreteness of both meronym and holonym candidates. The concreteness of a single concept is defined as the fraction of ontological sorts for a concept that are concrete (ontological sort: co or subordinated to co). Note that a concept can be assigned several ontological sorts if it is a meaning molecule (Helbig, 2006). A meaning molecule is assigned several meaning facets where each facet can have different ontological sorts, features and types of extensionality. Let us consider an example for the calculation of the concreteness. If a concept is assigned two concrete sorts and one abstract, then the concreteness of this concept is 2/3.

Ontological Sorts: Consider the case that at least one of the compared concepts involved in the meronymy hypothesis is a meaning molecule and is associated to several ontological sorts. The filtering process as described in Section 6.1 tests if there is at least one admissible combination of ontological sorts of the two concepts. All other ontological sorts are disregarded for this test. However, the disregarded ontological sorts can also give a clue about the hypothesis correctness. Therefore, we introduced a feature which employs the total set of ontological sorts of the two compared concepts.

A meronymy relation is presumably more likely to hold if the involved concepts are assigned similar sets of ontological sorts. Thus, this feature is set to the Jaccard coefficient considering the ontological sorts of the meronym (m) and holonym (h) candidates:

$$sort\_feature(m,h) := \frac{|sorts(m) \cap sorts(h)|}{|sorts(m) \cup sorts(h)|} \tag{13}$$

7 Meronymy Subrelations

The algorithm described thus far only extracts relations of type MERO. However, the meronymy relation is divided into several subrelations. In this section we describe how the correct subrelation is chosen. First, we introduce the set of all subrelations to choose from. Second the decision procedure is described in detail.

7.1 Types of Meronymy Relations

Winston proposes six subrelations for meronymy (Winston et al., 1987), which unfortunately are not sufficiently clearly defined. Therefore we base our approach on the meronymy relations of MultiNet (see (Helbig, 2006), Chapt. 4.2 and 18.2.49) which are systematically described and underpinned by an axiomatic apparatus. In the following it is tried to establish an approximate correspondence between both systems. Please note, that the division of meronymy of WordNet (Fellbaum, 1998) into three subrelations is considered by both aforementioned authors as being underspecified and not sufficiently differentiated. The following subrelations are proposed by Winston:

- Component-integral: A relation between an object and one of its components. Important for this relation is the fact that object and component can be perceived

separately from each other. MultiNet relation: PARS (see Section 3). Example: *A car wheel is part of a car.*

- Member-collection: This relation represents the membership in a set. MultiNet relation: ELMT. Example: *A soccer player is a member of a soccer team.*
- Portion-mass: Relations which refer to mass units and their parts. MultiNet relation: PARS, for temporal units: TEMP. Example: *A meter is part of a kilometer, a slice of the cake is part of the cake.*
- Stuff-object: This relation represents the chemical composition of an object. MultiNet relation: ORIGM^{-1} if the holonym denotes a physical object, otherwise PARS. Example: *Alcohol is part of wine. Steel is part of a bike.*
- Feature-activity: Activities can usually be divided into several subactions. MultiNet relation: HSIT. Example: The following subactions belong to the activity *going out for dinner*: *visiting a restaurant*, *ordering*, *eating* and *payment*
- Place-area: This relation holds between two objects if one of these objects is geographically part of the other object. MultiNet relation: PARS. Example: *Germany is part of Europe.*

Additionally, (Helbig, 2006) defines a further meronymy subrelation for subsets called SUBM in MultiNet. Example: *A brigade is a subset of a division.* Note that brigade/division is not a member-collection relationship since both, a division and a brigade, denote concepts with sets as their extension, whose elements are soldiers.

Premise	Decision
$etype(m) + 1 = etype(h)$	ELMT
$sort(m) \sqsubseteq si \wedge$ $sort(h) \sqsubseteq si$	HSIT
$sort(m) \sqsubseteq s \wedge$ $sort(h) \sqsubseteq d$	ORIGM^{-1}
$etype(m) = etype(h) \wedge$ $etype(m) > 0$	SUBM
$sort(m) \sqsubseteq ta \wedge$ $sort(h) \sqsubseteq ta$	TEMP
otherwise	PARS

Table 5: Selecting the correct meronymy subrelation (d=discrete object, s=substance, si=situation, ta=temporal abstracta, m=meronym, h=holonym), $sort(x) \sqsubseteq y :\Leftrightarrow sort(x) = y$ or $sort(x)$ is a subsort of y.

7.2 Selecting the Correct Meronymy Subrelation

The semantic lexicon is employed to choose the correct subrelation for an extracted meronymy relation, i.e., this process is not based on any machine learning algorithm. The decision rules are given in Table 5. For the definition of ontological sorts and the type of extensionality, which are required for the decision process, see Section 3. This procedure requires that both concepts taken into consideration are contained in the lexicon. If this is not the case, then one of the following fall-back strategies are used. First, if one of the regarded concepts is represented by a compound word, as determined by a morphological compound analysis, then the lexical entry of the base concept (concept corresponding to the correct reading of the base word) can be used. Second, for correct meronymy relations, which is assumed for this selection, the semantic sorts of the two concepts must usually be identical (exception: ORIGM^{-1}). This means, for instance, that if the first concept is known to be of sort *ta* the second should have the same sort and the correct subrelation should be TEMP. If a concept is a meaning molecule, the facets are chosen from both concepts for comparison, which are most similar in regard to ontological sorts and semantic features.

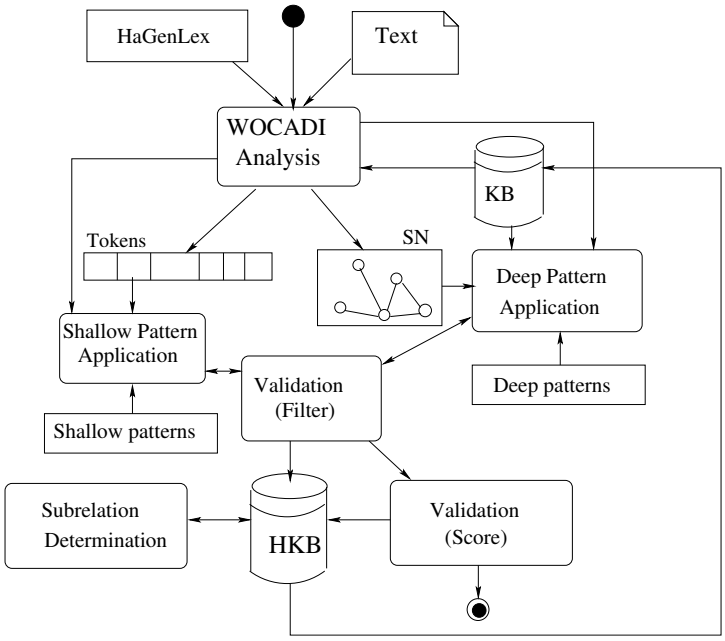

Figure 3: Activity diagram of the meronymy extraction done by SemQuire.

8 Architecture

To find meronymy relations from a text, this text is processed by our knowledge acquisition tool called SemQuire[8] (see Figure 3).

1. At first, the sentences of Wikipedia are analyzed by the deep linguistic parser WOCADI employing the knowledge base KB, containing the general background knowledge and the axiomatic apparatus.
   As a result of the parsing, a token list, a set of syntactic dependency trees, and a large semantic network (SN) are created.
2. Shallow patterns, consisting of a regular expression in the premise, are applied to the token lists, and deep patterns are applied to the SNs to generate proposals for meronymy relations (see Section 4 and Section 5).
3. A validation tool using ontological sorts and semantic features checks whether the proposals are at all technically admissible to reduce the amount of data stored in the hypotheses knowledge base HKB (see Section 6.1).
4. If the validation is successful, the meronymy candidate pair is added to the HKB. Steps 2–4 are repeated until all sentences are processed.
5. Each meronymy candidate pair in HKB is assigned a confidence score (see Section 6.2) estimating the likelihood of its correctness.
6. The correct meronymy subrelation is determined (see Section 7).
7. The highest scored hypotheses in HKB are manually inspected and eventually added to the knowledge base KB.

[8] SemQuire is derived from *acquire knowledge semantic-based*.

Feature	Correlation
Deep/Shallow	0.512
Ontological Sorts	0.436
Correctness Rate	0.397
Use of a Taxonomy	0.379
Concrete/Abstract	0.337
Mutual Information	0.030

Table 6: Correlation of features to relation correctness

9 Evaluation

The meronymy relations automatically acquired stem from the German Wikipedia corpus from November 2006 consisting of 500 000 articles and 20 million sentences.

In more than 6 million cases, a relation candidate was filtered out as being incorrect by our validation component. In total, 1 449 406 (different) meronymy relation hypotheses were finally stored in the knowledge base, 286 008 of them originating exclusively from deep patterns. In total, the relations in the knowledge base were extracted by approximately 2.5 million pattern applications.

	GermaNet		Costello		SemQuire		Sum
	PNM	PM	PNM	PM	PNM	PM	(PM+ PNM)
NM	750	0	506	244	666	84	750
M	718	32	393	357	151	599	750
Sum	1468	32	899	601	817	683	

Table 7: Confusion matrix for SVM optimization. NM = no meronymy relation present, M = meronymy relation present, PNM = predicted non-meronymy relation, PM = predicted meronymy relation.

For relation hypotheses extracted by deep and shallow patterns together, the precision is more than three times higher than for the relations that were only extracted by shallow patterns. 1 450 of the relations of the knowledge base were extracted alone by employing the SUB0 transitivity axiom A_2, exploiting the fact that a *district* de-

Measure	GermaNet	Costello	SemQuire
Accuracy	0.521	0.573	0.843
Recall	0.043	0.476	0.799
Precision	1.000	0.594	0.877
F-measure	0.082	0.528	0.837

Table 8: Accuracy, recall, precision, and F-measure for SVM optimization.

notes a *part* (see the example in Section 5). In total, logical axioms have been applied in 58 101 relation extraction processes, discovering 34 114 distinct relations. Table 4 shows a selection of axioms and the number of extracted hypotheses applying a certain axiom.

1 500 hypotheses were selected for the evaluation and annotated for correctness. Additional 50 000 hypotheses were annotated with their correctness exploited by the feature *Correctness Rate*, which is described in Section 6.2. Table 6 shows several scoring features and their associated correlation to the hypothesis correctness as specified by the annotators (one (1) for hypothesis is correct, zero (0) for incorrect).

Accuracy, precision, recall, F-measure, and confusion matrices were determined by a 10-fold cross-validation. Precision is the relative frequency with which a predicted meronym is actually one, while accuracy denotes the relative frequency with which the decision (meronymy/non-meronymy) is in fact correct. Note that recall does not relate to the extraction process, but rather only to the (score-based) validation. Thus, it specifies the relative frequency with which a correct relation in the data set of 1500 relations is actually identified as being correct by our system.

Our approach is compared with a GermaNet classifier as a baseline that predicts a relation of our hypotheses set to be meronymic if this relation is contained in GermaNet (GermaNet synsets are mapped to HaGenLex concepts semi-automatically) or can be derived by other GermaNet relations. A second baseline is the validation feature of Costello (Costello, 2007) which is also used as a feature by our system. The evaluation results are given in Tables 7 and 8. The evaluation showed that we were able to find a lot of meronyms not contained in GermaNet. In particular, less than 6% of the meronyms identified by SemQuire were contained or derivable from GermaNet. In addition, our extracted meronymy relations are, in contrast to GermaNet, all concept-based and not synset-based.[9]

They are also further differentiated into several subrelations. The correct subrelation was determined in 92.5% of the cases.

The results of our system are quite competitive in comparison with the results obtained by Girju et al. (2006) (precision: 0.81 and recall: 0.759). For this comparison, one has to take into consideration that Girju's approach is operated with English data and employs semantic relations from WordNet. Therefore, for his results, larger lexical resources were available than for German, which makes this task for a German text corpus more difficult.

The runtime of the algorithm heavily depends on the theorem prover timeout, i.e., the maximum amount of time which is available for a single proof. Currently the timeout is set to 0.1 seconds. With that the total runtime of the meronymy extraction algorithm is about three weeks on a Intel Core 2 Quad Q9550 CPU with 2.83 GHz using 8MiB of memory. By reducing the timeout the calculation time can be arbitrarily

[9] The difference is seen in the fact that MultiNet concepts are embedded in a complex linguistic and logical apparatus. Thus, concept ids of MultiNet are present in meaning postulates and other logical axioms, they are contained in the analysis results derived automatically by word disambiguation from the lexicon, and so on. This embedding in a whole process of language understanding is lacking in the case of representatives of synsets.

reduced. Naturally, the number of extracted hypotheses will decrease with descending timeout threshold. The realistic lower limit for the entire processing time, where still a reasonable amount of hypotheses can be found, is two days.

10 Conclusion and Future Work

In this paper, a logic-oriented approach has been presented for extracting meronymy relations from Wikipedia via text mining, which has proven its value in acquiring large stocks of knowledge. Unlike other approaches, our methods are based on a deep semantic representation, employing logical axioms. The use of axioms improves the generality of the method and can therefore increase the recall of the patterns in terms of the number of extracted meronymy relations. Furthermore, axioms can also be used to improve the precision of the patterns.

For future work, we plan on increasing the number of axioms and transferring this approach to other semantic relations.

The application of semantic patterns and axioms proved to be important for meronymy detection and is, in our opinion, an important step towards the use of real text understanding for future knowledge extraction systems.

Acknowledgement

We wish to thank all members of our department for their support. Especially we wish to thank Tiansi Dong for proof-reading this paper. This work is partly funded by the DFG project *Semantische Duplikatserkennung mithilfe von Textual Entailment* (HE 2847/11-1).

References

Baumgartner, P., Furbach, U., and Pelzer, B. (2007). Hyper tableaux with equality. In *Automated Deduction – CADE-21*, volume 4603 of *LNCS*, pages 492–507. Springer, Heidelberg, Germany.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 57–64, College Park, Maryland.

Bunescu, R. C. and Mooney, R. J. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 724–731, Vancouver, Canada.

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines.* Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Costello, F. J. (2007). UCD-FC: Deducing semantic relations using WordNet senses that occur frequently in a database of noun-noun compounds. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval)*, pages 370–373, Prague, Czech Republic.

Culotta, A. and Sorenson, J. (2004). Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 423–429, Barcelona, Spain.

Fellbaum, C., editor (1998). *WordNet An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

Friedmann, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29:131–163.

Gärtner, T., Flach, P., and Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. In *Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop*, pages 129–143, Washington, District of Columbia.

Girju, R., Badulescu, A., and Moldovan, D. (2006). Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.

Glöckner, I. (2007). Answer validation through robust logical inference. In Peters, C. et al., editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, number 4730 in Lecture Notes in Computer Science (LNCS), pages 518–521. Springer, Heidelberg, Germany.

Hamp, B. and Feldweg, H. (1997). GermaNet - a lexical-semantic net for german. In *Proceedings of the ACL/EACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain.

Hartrumpf, S. (2002). *Hybrid Disambiguation in Natural Language Analysis*. PhD thesis, FernUniversität in Hagen, Fachbereich Informatik, Hagen, Germany.

Hartrumpf, S., Helbig, H., and Osswald, R. (2003). The semantically based computer lexicon HaGenLex – Structure and technological environment. *Traitement automatique des langues*, 44(2):81–105.

Helbig, H. (2006). *Knowledge Representation and the Semantics of Natural Language*. Springer, Heidelberg, Germany.

Lyons, J. (2002). *Linguistic Semantics - An introduction*. Cambridge University Press, Cambridge, UK.

Pantel, P. and Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the Conference on Computational Linguistics/Annual Meeting of the Association of Computational Linguistics (COLING/ACL)*, pages 113–120.

Pantel, P. and Ravichandran, D. (2004). Automatically labeling semantic classes. In *Proceedings of the Human Language Technology Conference and the Conference of the North American chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL)*, pages 321–328, Boston, Massachusetts.

Reichartz, F., Korte, H., and Paass, G. (2009). Dependency tree kernels for relation extraction from natural language text. In *Proceedings of the European Conference on Machine Learning and the Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pages 270–285, Bled, Slovenia.

Suchanek, F. N., Sozio, M., and Weikum, G. (2009). SOFIE: a self-organizing framework for information extraction. In *Proceedings of the 18th International Conference on the World Wide Web (WWW)*, pages 631–640.

vor der Brück, T. and Helbig, H. (2010). Validating meronymy hypotheses with support vector machines and graph kernels. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA)*, pages 243–250, Washington, District of Columbia.

vor der Brück, T. and Stenzhorn, H. (2010). Logical ontology validation using an automatic theorem prover. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI)*, pages 491–496, Lisbon, Portugal.

Winston, M. E. et al. (1987). A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417–444.

Zhao, S. and Grishman, R. (2005). Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 419–426.

Frank Richter, Fabienne Fritzinger, Marion Weller

Who Can See the Forest for the Trees? Extracting Multiword Negative Polarity Items from Dependency-Parsed Text

1 Introduction

Ever since the groundbreaking work by Fauconnier (1975) and Ladusaw (1980), research on negative polarity items (NPIs) has been dominated by two fundamental assumptions about the licensing contexts of NPIs and their inherent semantic-pragmatic properties. The contexts in which NPIs may occur felicitously are said to have the semantic property of being *downward entailing* (which we will briefly explain below), and the elements themselves are often said to be located at the end of a pragmatically motivated scale, typically signalling a minimal amount, a smallest size, or similar concept. While the Ladusaw-Fauconnier theory has been substantially refined over time, and while there are very diverse variations on how the technical details of the theory are spelled out, its core insights are currently widely accepted and remain a point of reference for practically any 'formal' theory of NPIs. Some theories are syntactic in nature and formulate the relevant scope constraints relative to (possibly quite abstract) syntactic configurations, others are semantic and define hierarchies of negations of varying strength, and yet another group of theories is predominantly pragmatic, relying heavily on scalar implicatures, domain widening, and related concepts. There are, of course, also approaches in which syntax, semantics, and pragmatics all play a role. Overall, the number of papers and books that have been published on the subject of NPIs over the last 40 years is nothing short of intimidating.[1]

Given the sheer volume of the NPI literature, it is all the more surprising and striking that much of the discussion revolves around a very small set of items. Especially some of the most sophisticated and influential papers, such as Kadmon and Landman (1993), Krifka (1995), and Chierchia (2006), discuss hardly more than a handful of items, and some studies almost exclusively focus on one, *viz.* English *any*, which can be regarded as *the* classical example for a minimizer, with its variants *anything, anyone, anybody, anywhere*, etc. Since with *any* one of the most prominent items of interest is a minimizer, investigations into the significance of this particular property for the entire class of NPIs have turned into a dominating topic and occasionally even push aside the observation that being a minimizer is not a necessary (nor a sufficient) property of NPIs. As a result of its narrow empirical focus, the tendency to build a very comprehensive theory on an extremely small, carefully chosen but deeply researched set of examples is characteristic for large parts of the literature on NPIs. This might mean that only a fraction of the properties and behavior of NPIs are treated in current theories.

[1] To get a first impression of the amount of work published on the topic, the electronic NPI bibliography at www.sfb441.uni-tuebingen.de/a5/pib/XML2HTML/list.html is a good starting point. It lists well over 130 articles and books.

A more comprehensive overview of the landscape of empirical phenomena beyond the typical core examples of semantic and pragmatic studies of NPIs can be obtained by turning to research which approaches NPIs from a different angle. For German, Kürschner (1983) contains a collection of 344 (single-word and multiword) items which show strong affinity to negation and negative environments. Unfortunately, Kürschner's collection does not attempt a syntactic or semantic analysis within one of the major formal linguistic frameworks, and its data do not receive the kind of theoretical classification that would make them readily accessible to proponents of the Ladusaw-Fauconnier school. The most serious shortcoming in this respect might be the omission of a theoretically motivated distinction between items that strictly require licensing negation and those which do not grammaticalize this requirement and show only a preference for negative environments.

Another rich source for a broader picture of the empirical facts is provided in the work of a Dutch group of linguists around Jack Hoeksema, Ton van der Wouden, and Frans Zwarts. In contrast to Kürschner's strongly data-driven compilation, Jack Hoeksema's comprehensive studies of (predominantly Dutch) NPIs combines theoretical concepts from the Ladusaw-Fauconnier tradition with extensive synchronic and diachronic corpus studies. Hoeksema (1997) is an early example of the potential of using electronic corpora in this area and gives a first glimpse of the highly differentiated and exciting landscape of polarity phenomena that emerges when corpus data is systematically researched and investigated with the tools of formal NPI research.

Kürschner's collection and the comprehensive work of the above group of Dutch linguists on a wide range of polarity phenomena inspired the creation of a collection of theoretically classified and richly documented German NPIs as a subcollection of the electronic *Collection of Distributionally Idiosyncratic Items* (CoDII, Trawiński et al. (2008); Richter et al. (2010)).[2] The NPI collection in CoDII can be considered a thorough inventory of our current knowledge of the extent of German NPIs. Its empirical coverage subsumes all available sources, i.e. it lists all German NPIs mentioned in all of the literature that was surveyed in its creation, including the true NPIs in Kürschner's list. But CoDII not only collects items discussed elsewhere and reports their usage in systematically selected licensing environments with naturally occurring examples from corpora. Its compilation was also supported by search results from the implementation of the first semi-automatic extraction procedure for NPIs from corpora (Lichte, 2005a,b)[3].

The collection of German NPIs in CoDII and the NPI extraction procedure of Lichte and Soehn (2007) form the starting point of the present study. The motivation behind the new NPI extraction procedure which we will present is to prepare a wider empirical base for a future, more comprehensive theory of NPIs, to sharpen our understanding of the syntactic and semantic diversity of NPIs, and to provide the necessary material for psycholinguistic studies and the use of NPIs in language processing tasks. Our

[2] CoDII was compiled in a project of the former *Sonderforschungsbereich 441* and is available at www.sfb441.uni-tuebingen.de/a5/codii/

[3] Subsequently refined in Lichte and Soehn (2007)

immediate objective is to demonstrate that our method can significantly extend the set of known NPIs in German (as represented by the 165 entries in CoDII, the largest collection available today). In the absence of a complete repository of German NPIs that could serve as a gold standard, we will measure the success of our method by the number of items we can add to the CoDII collection.

Due to the striking frequency of multiword NPIs in CoDII, and based on the assumption that there is an affinity between the properties of NPIs and at least some classes of idiomatic expressions, our new method targets multiword NPI candidates. We adapt an extraction pipeline that was previously successfully applied in the identification of multiword expressions (MWEs, (Fritzinger and Heid, 2009)) using statistical association measures and two linguistically motivated scores, the degree of morpho-syntactic fixedness (Weller and Heid, 2010) and semantic opacity (Fritzinger, 2009) of an expression. The significant difference between our method and the basic form of the earlier algorithm by Lichte and Soehn is our focus on MWEs. Lichte and Soehn primarily search for single-word NPIs and capture multiword NPIs only indirectly in an extension to their basic extraction mechanism by building lemma chains of length $n+1$ from lemma chains of length n and checking if extending a lemma chain makes it a better NPI candidate.

Section 2 gives a very brief overview of NPIs and their licensing contexts. In Section 3 we characterize our corpora and our extraction method for MWE candidates, before we say more about how we model NPI licensing contexts in Section 4. In Section 5 we present optimizations to the statistical processing of NPI candidates that we apply to achieve a higher ratio of NPIs at the top of our candidate lists, and we propose some linguistic measures for the identification of idiomatic candidate expressions. Section 6 discusses the results. We conclude with a short outlook on future work in Section 7. The appendix lists the NPIs that our extraction method found.

2 npis and npi Licensing

NPIs are defined as single words or multiword expressions which require the presence of an appropriately 'negative' element in their utterance context. The negative element is said to *license* the NPI, and without a proper licenser the presence of an NPI results in ungrammaticality. Examples of extensively researched NPIs from English are the determiner *any*, the adverb *ever*, and the MWEs *red cent* and *to lift a finger*; good licensers are the sentential negation adverb *not* or negative quantifiers such as *no students*. In (1a/b)–(4a/b) we see sentence pairs with the NPI *any* which is licensed here in the scope of four different lexical licensers ((a)-sentences; licensers are underlined). The sentences become ungrammatical when the licenser is omitted or replaced by an element without the necessary licensing properties ((b)-sentences). The (c) and (d) sentences are parallel German counterparts to the standard English examples in (a) and (b), with the verb *scheren* ('to care') as NPI.

(1) a. Pat did <u>not</u> see **any** student in the hallway this morning.
b. *Pat saw **any** student in the hallway this morning.

c. Peter **schert** sich nicht um Lokalpolitik.
Peter cares REFL not about local politics
'Peter does not care about local politics.'

d. *Peter **schert** sich um Lokalpolitik.
Peter cares REFL about local politics
'Peter cares about local politics.'

(2) a. Nobody saw **any** student in the hallway this morning.

b. *Everybody saw **any** student in the hallway this morning.

c. Niemand **schert** sich um Lokalpolitik.
nobody cares REFL about local politics
'Nobody cares about local politics.'

d. *Jeder **schert** sich um Lokalpolitik.
everybody cares REFL about local politics
'Everybody cares about local politics.'

(3) a. Kim never saw **any** student in the hallway.

b. *Kim saw **any** student in the hallway this morning.

c. Peter **schert** sich niemals um Lokalpolitik.
Peter cares REFL never about local politics
'Peter never cares about local politics.'

d. *Peter **schert** sich um Lokalpolitik.
Peter cares REFL about local politics
'Peter cares about local politics.'

(4) a. Few lecturers saw **any** student in the hallway this morning.

b. *Some lecturers saw **any** student in the hallway this morning.

c. Wenige Bundespolitiker **scheren** sich um Lokalpolitik.
few federal politicians care REFL about local politics
'Few federal politicians care about local politics.'

d. *Einige Bundespolitiker **scheren** sich um Lokalpolitik.
some federal politicians care REFL about local politics
'Some federal politicians care about local politics.'

The question of how to characterize the necessary negativity more accurately and which structural, logical or pragmatic relationship must hold between an NPI and its licenser or licensing environment has been subject to intense debate in theoretical linguistics, and is far from being settled. According to the dominant view, the contextually necessary negativity can best be semantically characterized in terms of the entailment behavior of the licensing environment, and the entailment behavior is triggered by an

operator that must stand in a certain structural relation to the licensed NPI. NPIs are licensed in the semantic scope of the relevant operators, and are ungrammatical in their absence (see Zwarts (1997) and van der Wouden (1997) for details). Note that a component of a larger constituent can be responsible for the licensing behavior of that constituent. The NP quantifier *few lecturers* in (4a) is a licenser due to the logical behavior of its determiner, *few*, as can be verified by the ungrammaticality of (4b), where *few* has been replaced by the determiner *some*.

To keep our terminology simple, in the remainder of this paper we will call all relevant licensing environments *negative*. It is important to keep in mind that, despite this naming convention, other operators whose negativity is much less apparent than in the case of sentential negation and negative quantifiers can also license NPIs. Examples of weaker forms of negation are the quantifier *few lecturers* (see (4a)) and questions, which are perfectly valid licensers for many NPIs. Most licensing environments are logically *downward entailing*, which means that they allow inferences from supersets to subsets. For example, the downward entailing operator *few doctors* is responsible for the valid inference from the truth of *few doctors recommended showers* to *few doctors recommended cold showers*. Questions are sometimes subsumed under a weaker class of negativity, called *nonveridicality* (Zwarts, 1995). Nonveridical operators prohibit inferring the truth of a proposition from it being uttered: *Did Peter come late?* does not entail that Peter came late.

The examples in (5)–(8) illustrate multiword NPIs and highlight additional factors that need to be taken into consideration when searching for them in corpora and when checking if a candidate expression is indeed an NPI. English examples are followed by their German translations into corresponding constructions with NPIs. All explanations below about the English examples also apply, *mutatis mutandis*, to their German counterparts.

(5) a. John didn't **drink a drop** (of alcohol) last night.
b. #John **drank a drop** (of alcohol) last night.
c. #Few students **drank a drop** (of alcohol) last night.
d. Hans hat letzte Nacht keinen **Tropfen** (Alkohol) **getrunken**.
e. #Hans hat letzte Nacht **einen Tropfen** (Alkohol) **getrunken**.
f. #Wenige Studenten haben letzte Nacht **einen Tropfen** (Alkohol) **getrunken**.

(6) a. Nobody had **the slightest inkling** about where to go.
b. *Few visitors had **the slightest inkling** about where to go.
c. Niemand hatte **die leiseste Vorstellung**, wohin man gehen sollte.
d. *Wenige Besucher hatten **die leiseste Vorstellung**, wohin sie gehen sollten.

(7) a. Thomas isn't **much of a** soccer player.

b. *Miroslav is **much of a** soccer player.

c. Thomas ist **beileibe** kein Fußballer.

d. *Miroslav ist **beileibe** ein Fußballer.

(8) a. This sentence will not parse **in a million years**.

b. #This sentence will parse **in a million years**.

c. Dieser Satz lässt sich **im Lebtag** nicht parsen.

d. *Dieser Satz lässt sich **im Lebtag** parsen.

A comparison between (5a) and (5b) shows that with multiword NPIs it becomes important to distinguish between different readings of candidate expressions. In (5a) the idiomatic reading (John did not drink any alcohol at all) is very prominent, whereas (5b), due to the absence of an appropriate licenser for the idiomatic NPI *drink a drop*, does not have the idiomatic reading. However, there is a literal meaning (the amount that John drank was one drop), which is in principle available. In (5) and (8) we indicate unavailable idiomatic NPI readings and available literal readings with '#'. In cases in which there is no literal meaning ((6) and (7)) this complication does not arise. The examples also demonstrate that (simple and complex) NPIs can be of almost any syntactic category. The present selection of multiword NPIs also provides examples for the class of minimizers (and maximizers, (8a)), which play such a prominent role in current pragmatic theories of NPIs. Finally, (5c) shows that the licensing requirements of NPIs may differ: (4a) confirmed that quantifiers with the determiner *few* license *any*, but *drink a drop* is not licensed by a corresponding quantifier in (5c), it needs a *stronger* type of negation such as sentential negation to be satisfied (5a). Throughout this study, we will ignore the observation that licensing requirements can be of different strength.[4]

For the present research, we follow an idea applied in the NPI extraction algorithm by Lichte and Soehn (2007) and exploit the fact that a finite set of particular lexemes (determiners, adverbs, subordinating conjunctions, a small number of verbs) and an equally small set of syntactic structures (antecedents of conditionals, questions, comparative constructions) are good indicators of licensing environments.[5] Although they do not cover all possible licensing environments, and although there can be additional syntactic or semantic properties present in a clause which prevent NPIs from being licensed in certain positions, we assume that we can detect enough licensing environments sufficiently well to obtain useful NPI candidate lists when using our heuristics in large corpora.

[4] See Lichte and Soehn (2007) for an attempt to use a hierarchy of three types of negation strengths in extracting NPIs from corpora.

[5] Details about this choice of licensing contexts (which is derived from the data-driven classification of NPIs in CoDII) will be discussed in Section 4, along with an explanation of why there is an unavoidable residue of licensing environments which cannot be detected with our type of heuristics. In Section 5.3 we illustrate concrete limitations of our extraction pipeline with problematic data from our corpus.

3 Preliminaries: Extraction Methodology

Finding multiword NPIs in corpora is not an easy task: NPIs are known to be rare, and many members of the subclass of multiword NPIs are probably even rarer. Lichte and Soehn (2007) report that they found 28 occurrences of the single-word NPI *Menschenseele* ('living soul'), and the same number of occurrences of the complex NPI *alle Tassen im Schrank haben* ('to have lost one's marbles'). In EUROPARL (see below) we found 36 occurrences of *ein Hehl aus etwas machen* ('to hide sth.'), 18 occurrences of *ein Blatt vor den Mund nehmen* ('to mince words'), and 6 occurrences of *einen Finger rühren* ('to lift a finger'). It is evident that in order to retrieve enough occurrences of an expression to apply statistical methods which can meaningfully support its association with negative environments, we would thus like to use as large a corpus as possible. Easily available unannotated text would fulfill this desideratum.

At the same time we are faced with a second requirement. Detecting negative environments and determining that several words form a multiword expression presupposes linguistic knowledge. For that reason these two tasks can be most easily accomplished with text that is already linguistically annotated and provides a syntactic basis for recognizing plausible multiword expression candidates and at least some indication of the scope of relevant semantic operators.[6] If we hypothesized naively, i.e. without paying attention to structure, that any collection of words in a sentence could be an MWE candidate, we would quickly run into an intractable combinatorial explosion of candidates. Syntactic knowledge about which groups of words form meaningful syntactic units is particularly relevant for languages with discontinuous constituents such as German. In short, we can benefit from annotation. Annotated corpora are, however, limited in size, and decrease the size of the available data base compared to plain text.

Our method tries to strike a balance between the conflicting needs of working with a large resource and being able to refer to structural linguistic knowledge. As outlined in Section 3.1, we start from unannotated corpora, and we obtain the necessary annotation by relying on a robust broad coverage dependency parser with rich lexical information. Our next step then, described in Section 3.2, is to extract certain complex syntactic patterns that we consider promising structural skeletons of multiword NPIs. The MWE candidates that we extract in this preprocessing step will later provide the foundation to finding those complex expressions that are statistically associated with NPI licensing contexts.

3.1 Data

NPIs are infrequent in text corpora and, presumably, in everyday language. In order to avoid problems in the statistical analysis arising from sparse data, we need a very large text corpus to start from.[7] An overview of our corpus collection is given in Table 1.

[6] Lichte and Soehn's work was based on a treebank newspaper corpus. Their heuristics for determining the scope of semantic operators was using the syntactic structure provided by the tree annotation.

[7] Lichte (2005b) reports that he achieves the best UAP index (see Section 5.2) for his candidate lists with a minimal frequency of 60 for the items considered in the candidate list. Since this threshold

It contains about 269 million words (tokens), comprising text from several German newspapers, and the proceedings of the European parliament debates, EUROPARL (Koehn, 2005).

source	size (tokens)	text type	years
Europarl	35 million	debates	1996-2006
Frankfurter Allg. Zeitung	70 million	news	1997-1998
Frankfurter Rundschau	40 million	news	1992-1993
Handelsblatt	36 million	news	1986/1988
Stuttgarter Zeitung	36 million	news	1991-1993
Die Zeit	52 million	news	1995-2001
Total:	269 million		

Table 1: Composition of the dataset

EUROPARL will play a distinguished role in our NPI identification procedure when we target semantic properties of MWEs. At the beginning we will not do this yet and will only use the German part for monolingual processing. In later refinements of our method in which more linguistic knowledge is brought to bear, we also add the English, French and Swedish parts for multilingual processing. These refinements will be discussed in Section 5.2 below. For the initial identification of MWE candidates, we rely entirely on monolingual processing.

3.2 Multiword Extraction with Syntactic Patterns

In German, the constituent words of multiword constructions are not always adjacent to each other for two reasons. The possibility of constituent order variations in the middle field entails that multiword expressions may be realized discontinuously, with other constituents potentially intervening. The additional alternation of the verb between a verb second and verb final position in finite sentences means that a finite verb may precede or succeed those of its dependents that occur in the middle field. The sentence in (9) contains the NPI *(k)einen blassen Schimmer haben* (lit. '(not) to have a pale gleam'; '(not) to have the faintest idea'). The finite verb form *hat* in verb second position is linearly separated from its accusative object *blassen Schimmer* with which it forms a multiword NPI. Note that in a corresponding verb final construction *blassen Schimmer* would precede *hat*, thus reversing their order, and they could be adjacent.

(9) Er **hat** zum jetzigen Zeitpunkt keinen **blassen Schimmer**...
he has at this point no pale gleam...

excludes too many NPIs that enter the candidate lists with a lower choice of minimal frequency, Lichte (2005b) decides to choose a minimal frequency of 40, despite the ensuing increase of noise in the candidate list; Lichte (2005a) lowers the threshold even further (to 20), whereas Lichte and Soehn (2007) chooses 30.

'At this point he does not have the faintest idea.'

Deep syntactic analysis is essential in order to reliably extract potentially discontinuous multiword constructions. In the past, we successfully used the dependency parser FSPAR (Schiehlen, 2003) for a variety of MWE extraction tasks. FSPAR is highly efficient and relies on a large lexicon. It processes about 10 million words in 30 minutes, is very robust and includes enough morphological information for our task (see Figure 1a, 5th column), which means that we do not need a separate morphological analyzer. FSPAR is compatible with the German orthographical conventions before and after the spelling reform of 1996. An example analysis of FSPAR is given in Figure 1. It shows a dependency structure for the sentence *Und er hat keinen blassen Schimmer, was gerade vor sich geht* ('And he doesn't have the faintest idea what is going on').

The dependency tree representation in Figure 1b is not provided by the parser in this format (but can be inferred from its analysis); we insert it here in order to enhance the readability of the example. The original FSPAR output given in Figure 1a contains the following information (from left to right): position of the token in the sentence, token, part of speech tag, lemma, morpho-syntactic information, dependency relation (the numbers refer to sentence positions in the first row), and grammatical function.

The dependency parses provide all necessary information for building a collection of multiword items that we can investigate for their distributional properties. Those multiword items that occur in NPI licensing environments will become our NPI candidates. To obtain syntactically meaningful units we first identify in the corpus certain patterns of verbs and their dependents. The patterns we collect are verbs and dependents that are nouns, adjective-noun combinations, preposition-noun combinations, preposition-adjective-noun combinations, or noun plus preposition-noun combinations. These patterns are chosen because the class of verbal MWEs and verb phrase idioms is known to be comparatively large, and we expect to find a sizable number of NPIs among them.

In order to extract the target patterns from the dependency analyses, we employ Perl scripts. Starting with all lexical verbs found in a sentence (such as *haben* in the example in Figure 1), these scripts collect the subject and objects (*Schimmer*), including modifying adjectives (*blassen*), and the prepositional phrases related to the initial verb. To accomplish this task, the extraction scripts refer to part of speech tags and morphological features, and to the dependency structure given in the second to last column of the FSPAR output. While no other information is needed to identify the dependency patterns of interest, we still gather additional syntactic features for later use. All accessible morpho-syntactic information including the type of determiners, syntactic number features, and the presence of comparative forms (for adjectives) is collected at this point already for linguistic post-processing at a later stage (Section 5.2).

The extracted candidate items, consisting of the lemmas of the verb, objects, the subject, and prepositional phrases, are stored together with the available morpho-syntactic information in a PostgreSQL database (see Weller and Heid (2010) for details). The database entry thus obtained for the verb+object pair *Schimmer haben* in Figure 1 is shown in Table 2.

(a) FSPAR output

0	Und	KON	und	|	2	ADJ
1	er	PPER	er	Nom:M:Sg	2	NP:1
2	hat	VVFIN	**haben**	3:Sg:Pres:Ind	-1	TOP
3	keinen	PIAT	kein	|	5	SPEC
4	blassen	ADJA	**blaß**	|	**5**	ADJ
5	Schimmer	NN	**Schimmer**	Akk:M:Sg	**2**	PCMP
6	,	$,	,	|	2	PUNCT
7	was	PRELS	was	Nom:N:Sg	11	NP:1
8	gerade	ADV	gerade	|	9|11	ADJ
9	vor	APPR	vor	Dat|Akk	11	ADJ
10	sich	PRF	er|sie|es	Dat|Akk	9	PCMP
11	geht	VVFIN	gehen	3:Sg:Pres:Ind*	2	ADJ
12	.	$.	.	|	-1	

(b) Tree representation

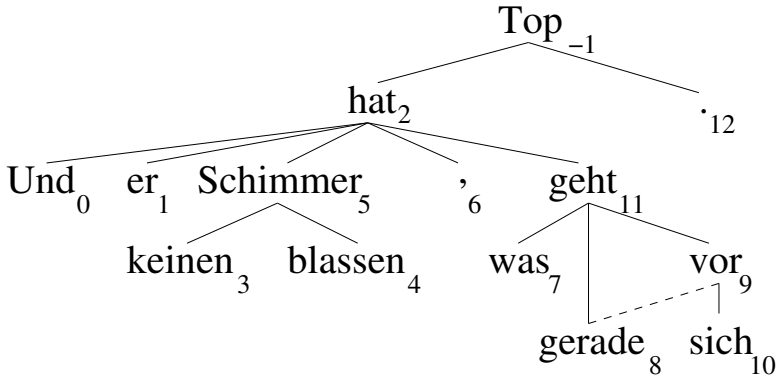

Figure 1: Dependency analysis of a sentence

The PostgreSQL database contains every dependency relation that the dependency parser makes available in its parse output and could be relevant to our NPI discovery procedure. With the database in hand it is now possible to work with patterns of varying form and length. In the present study, we choose to investigate five patterns: verb+object (NV), adjective+object+verb (ANV), preposition+noun+verb (PNV), noun+preposition+noun+verb (NPNV) and preposition+adjective+noun+verb (PANV). Examples for each of the patterns are shown in Table 3a; their occurrence frequencies can be seen in Table 3b. Since at this stage we have not yet done any checks regarding their occurrence inside or outside of NPI licensing contexts, the large majority of items are not NPIs. Most items are trivial combinations of words in the sense that they are not even MWEs but consist of free combinations of words that simply obey the general grammatical mechanisms of syntactic and semantic selection. Other items are statistical collocations with compositional semantics, and some are idiomatic expressions. Finally,

v_lem	subj	acc_obj	acc_obj_det	acc_obj_num	acc_obj_mod
haben	er	Schimmer	kein	sg	blaß

Table 2: Database entry for *Schimmer haben*

a few of our items are NPIs. In Table 3a we see one example of a compositionally constructed phrase ('trivial combination'), one of an idiomatic expression and one of an NPI for each of the five patterns. Due to the fact that some patterns are subsets of others (ANV obviously forms a subset of NV), their respective candidates occur in the results of their super-patterns as well (e.g. *Faden verlieren* ('lose one's train of thoughts'), is in the NV result list, while the complete expression, *roten Faden verlieren*, is contained in the results for ANV[8]). The table also reflects the fact that we extract lemmas instead of word forms.

(a) Examples for investigated patterns

pattern	trivial combination, idiomatic expression, NPI
NV	*Frau danken*, *Rede halten*, *Hehl machen*
ANV	*sachlich Grund sehen*, *rot Faden verlieren*, *blass Schimmer haben*
PNV	*auf Agenda stehen*, *unter Druck setzen*, *über Herz bringen*
NPNV	*Herr für Rede danken*, *Wind aus Segel nehmen*, *Blatt vor Mund nehmen*
PANV	*zu neu Debatte führen*, *für bar Münze nehmen*, *mit recht Ding zugehen*

(b) Occurrence frequencies of the patterns

	NV	ANV	PNV	NPNV	PANV
types	2 069 393	1 143 104	6 337 849	3 033 148	2 475 122
tokens	5 194 941	1 442 865	11 420 865	3 388 758	2 906 645

Table 3: Overview of syntactic patterns

Having completed the extraction of all expressions from the German corpus that meet our five pre-defined syntactic dependency patterns, everything is set up for the identification of syntactically complex NPIs. As complex NPIs are (possibly idiomatic) collocations, this step will ultimately filter out trivial combinations, leaving collocations and idiomatic phrases in negative environments.

Before we describe the statistical processing and linguistic refinements for targeting idiomatic NPIs in Section 5, Section 4 discusses how we identify NPI licensing environments. This is another nontrivial task, as we assume that NPI licensing is effected by

[8] This particular example is even more intricate, since *Faden verlieren* might be considered an independent idiomatic expression with a meaning that differs slightly from *roten Faden verlieren* ('losing the central idea'). In this case we could say that we did in fact find two idiomatic expressions.

a mixture of semantic, syntactic and pragmatic conditions which cannot be read off directly from the dependency information available in the extracted patterns.

4 Modelling Negative Contexts

Following the lead of Lichte (2005a) and Lichte and Soehn (2007), we identify the negative licensing contexts of multiword NPIs on the basis of certain syntactic configurations and a finite list of determiners, verbs, adverbs and other lexical elements. A few examples are listed in Figure 2.

sentential negation adverb	*nicht*
negative determiner	*kein*
nouns	*niemand, nichts*
adverbs	*kaum, nur, selten, wenig, ebensowenig, nie, niemals, nirgendwo, nirgends, nirgendwohin, nirgendwoher, keinesfalls, keineswegs*
inherently negative verbs	*ablehnen, anzweifeln, abstreiten, bestreiten, bezweifeln, dementieren, verhindern, verweigern, weigern*
negative conjunctions[9]	*ohne zu, ohne dass, ob, bevor*

Figure 2: A selection of lexical licensing contexts

Our extraction procedure comprises a component that recognizes negative contexts by the presence of at least one of our lexical or structural criteria for licensing contexts. Whenever an MWE occurs in such a context, that occurrence of the MWE is labelled with NEG, otherwise with NONEG. This format meets the requirements of the statistical association measures that are applied (Section 5) to distinguish multiword NPIs from other MWEs that might occasionally occur in a negative polar environment.

Let us consider four examples for licensing contexts we found for the NPI *alle Tassen im Schrank haben* (lit. 'to have all cups in the cupboard'; 'to have lost one's marbles'), which illustrate the wide variety of licensing possibilities found in corpora:

(10) Nicht **alle Tassen im Schrank** zu **haben** mag ja durchaus
not all cups in-the cupboard to have may PART indeed
produktiv sein für derlei Theater.
productive be for such theater
'Being somewhat insane may in fact be an advantage for this type of theater.'

[9] Recall that we use the word 'negative' loosely to designate environments which license NPIs. Here we mean to characterize subordinating conjunctions which license NPIs in the embedded clause.

(11) Kein Mörder, der **alle Tassen im Schrank hat**, würde mich
no murderer who all cups in-the cupboard has would me
umbringen.
kill
'No sane murderer would kill me.'

(12) ...sollte sich darüber hinaus allerdings fragen lassen, ob
...should himself moreover however ask let if
Vorstansdsmitglied P. S. noch **alle Tassen im Schrank hat**
board member P. S. still all cups in-the cupboard has
'...should seriously be wondering if board member P. S. has lost his marbles.'

(13) Jeder, der noch seine **fünf Tassen im Schrank hat**, weiß,
everybody who still his five cups in-the cupboard has knows
daß...
that...
'Any sane person knows that...'

In (10) the verb *haben* ('to have') is simply modified by the sentential negation adverb *nicht* ('not'), exemplifying the most straightforward case. Similarly, in (11), the subject noun phrase *kein Mörder* is a negative quantifier due to the determiner *kein* ('no'). In the construction in (12), the clause containing the expression *alle Tassen im Schrank haben* is an indirect question, which is a legitimate nonveridical licensing environment of the NPI. (13) is an instance of NPI licensing in the restrictor of a universal quantifier, in this case the nominal quantifier *jeder* ('everyone'). Restrictors of universal quantifiers are downward entailing, which is the most important semantic licensing condition. Replacing the universal with a proper noun or a definite noun phrase removes this semantic property and results in an ungrammatical utterance.[10]

The choice of lexical and structural indicators of negative environments for our extraction procedure is determined by two considerations: First, we use some of the lexical (and structural) licensers which CoDII lists. These elements and structural environments reflect the available linguistic knowledge in the NPI literature about licensing environments. Their practical usefulness for semi-automatic NPI extraction was confirmed by the results of Lichte and Soehn's work. Second, our choice of lexical and structural licensing environments is influenced by the type of structural information we expect to be available after running FSPAR. Here we follow our judgment about the reliability of the output of the dependency parser.

There are some obvious limitations to our selective and rather syntactic approach to modelling negative contexts. Since there are, in principle, infinitely many forms of valid licensing environments, it is impossible to define a syntactic pattern for every

[10] The substitution of *seine fünf* ('his five') for *alle* ('all') in the phrase *alle Tassen* in (13) is an instance of creative language use and does not change the (relevant aspects of the) meaning of the expression.

single one of them. The situation would become even more difficult if we decided to try to systematically detect cases in which a given pattern is not a licenser due to additional effects such as intervening quantifiers between a licenser and a potential licensee. This task would minimally presuppose some analysis of quantifier precedence conditions and would involve a closer investigation of the interplay between word order and syntactic structure. Moreover, some licensing environments are just not reliably identifiable without deep syntactic or semantic analysis. Examples in German are extraposed relative clauses (which might be in a downward entailing environment depending on the noun phrase they attach to), comparative clauses with adjectives plus *als*-clause (which require a reliable semantic analysis), subjunctive clauses, and opaque conditionals of the form *You say anything, and I kill you* (with *anything* being an NPI licensed by the conditional construction).[11] Our working assumption is that our model captures a sufficiently large portion of NPI licensing environments to produce good enough candidate lists. As long as we recognize enough actually occurring licensing environments and do not miss too many, and as long as the corpus is large enough, the statistical analysis should be able to cope with the noise caused by the lack of sophisticated semantic annotation.[12]

5 Optimization

At this point of our procedure, we have extracted a very large number of NPI candidates. The figures in Table 3 show that this is not a list that a human annotator could effectively work with to identify true NPIs. Amongst the items in the list are valid NPIs and other idiomatic multiword constructions, but the vast majority are trivial combinations of words, i.e. syntactically regular and semantically transparent constructions such as *auf Stuhl setzen* ('on chair sit'; 'to sit down on a chair'). Some of them might have an accidental high co-occurrence ratio with negative contexts in our corpus, and it is important to face the fact that there is no automatic procedure to validate NPIs. Manual annotation remains an indispensable step for our extraction method. A native speaker has to check if the use of a candidate expression without a negative context always leads to ungrammaticality. The question to decide is whether it is categorically impossible to use a candidate expression felicitously (under constant meaning) without a licensing context. Even strong statistical tendencies in large corpora cannot guarantee that this is the case for a given expression. Especially for idiomatic NPI candidates that permit a related literal meaning it can even be very hard for a native speaker to verify

[11] This list of difficult cases is taken from a slide presentation by Timm Lichte.

[12] As a reviewer succinctly remarks, our considerations here are *full of speculative assumptions*. In the presumed absence of an even remotely complete list of NPIs in German (or any other language), and confronted with a complete lack of the type of deep semantic analysis of large text collections that we would need to be able to identify all possible semantic licensing environments known from the linguistic literature, the only justification of our optimistic tone is the actual success of the method, measured by the number of new German NPIs that we find. There is much room for improvement.

that the idiomatic reading strictly requires a licensing context, because this fact might be concealed by the literal reading, which does not.

These complications aside, the key to success for semi-automatic NPI extraction is that some features that are characteristic for NPIs such as their significant co-occurrence with the licensing contexts described in Section 4, and the syntactic fixedness of idiomatic expressions can be automatically accessed. In the following sections, we describe how we used some of these features to create a list of manageable size with an enhanced number of valid NPIs at the top of the list by sorting candidates according to associative strength with their respective negative contexts and with linguistic features (morpho-syntactic fixedness or translational behavior). This preprocessing considerably reduces the necessary but time-consuming manual annotation efforts, and makes human annotation feasible.

5.1 Statistical Processing

A number of statistical association measures such as *log likelihood ratio* or *t-score* have been successfully applied to identify MWEs (see e.g. Evert (2004)). They indicate the associative strength of a word pair by taking into account the observed vs. expected frequencies of pairs and of their components in isolation. Assuming that NPIs are significantly associated with their negative context, we compute the associative strength between each MWE and its context label (which is NEG for negative contexts, and NONEG otherwise) to determine whether a negative context is obligatory for an expression. An example pair is: (*blassen::Schimmer::haben*, NEG).

We used the UCS toolkit[13] to calculate five standard association measures for each of our five candidate lists (cf. Table 3, with each candidate represented in lemma form). These lists were then sorted in decreasing order according to the resulting scores. Finally, the 500 highest scoring candidates with a strong statistical tendency to be associated with a NEG context label were manually annotated: '+' for valid NPIs and '−' for MWEs or trivial combinations. Since longer patterns are extensions of shorter patterns, there is a certain overlap between the items we find in longer and shorter patterns.

[13] UCS toolkit: www.collocations.de (Evert, 2004)

NPIs in top 500	NV	ANV	PNV	NPNV	PANV
log-likelihood	21	74	28	5	4
t-score	16	65	21	5	4
z-score	21	76	29	5	4
poisson	29	77	31	5	4
chi-squared	21	76	30	5	4

Table 4: NPIs found for each of the syntactic patterns when sorted according to standard association measures

The numbers of valid NPIs found amongst the top 500 candidates can be seen in Table 4. Even though *poisson* slightly outperforms the other measures, all results turned out to be quite similar. Furthermore, we also found that the NPIs were often the same: All 16 NPIs of the category NV found in the *t-score* sorting are a subset of those found by *log-likelihood*, *z-score* and *chi-squared*, while all 21 NPIs found by the latter ones are contained in the results for *poisson*. Similar observations were made for the other syntactic patterns.

(a)

	NPNV-pattern with negative context	f	position		
			POIS	LL	f
+	**Blatt vor Mund nehmen**	139	1	1	50
-	Angabe über Höhe machen	78	2	2	160
-	Richtlinie in Recht umsetzen	61	3	3	262
-	Ziel aus Auge verlieren	116	4	4	76
+	**Wald vor Baum sehen**	50	5	7	367
-	Angabe über Kaufpreis machen	42	6	6	466
(+)	**Hehl** aus Sympathie **machen**	38	7	8	561
(+)	**Hehl** aus Enttäuschung **machen**	37	8	9	594
-	Arbeit für Stunde niederlegen	37	9	11	573
(+)	Gefahr **von Hand weisen**	36	10	10	736
-	Stein in Weg legen	84	11	13	142
-	Zugang zu Trinkwasser haben	29	12	12	896
-	Änderungsantrag aus Grund akzeptieren	36	13	16	612
+	**Mördergrube aus Herz machen**	28	14	14	868
(+)	**Hehl** aus Abneigung **machen**	28	15	17	868

(b)

	PNV-pattern with negative context	f	pos (POISSON)	pos (f)
+	**aus Staunen herauskommen**	60	48	8998
+	**über Weg trauen**	91	51	6941
+	**mit Wimper zucken**	26	289	33412

Table 5: Samples of log-likelihood orderings for two patterns: (a) NPNV: poisson and log-likelihood and (b) PNV

Table 5a shows the top 15 entries of the NPNV pattern that are labelled with NEG. The candidates are ordered according to their *poisson* scores. The first column contains the manual annotation that reflects the judgment of the human annotators whether or not the expression is an NPI (+/−). The entries marked with '(+)' would be complete with only one noun, and therefore belong to the NV class rather than NPNV. Conversely, there

are patterns containing candidates that are not yet complete. The absolute frequency of the NPI candidates is indicated in column 3 (labeled *f*) while the last columns give the ranks of each expression according to different association measures (*poisson*, *log-likelihood* and *frequency* ordering, respectively). Note that the ranks obtained by the *poisson* method and *log-likelihood* do not differ substantially.

Since most NPIs are relatively infrequent, they would be hard to find in a list sorted by absolute frequency.[14] Sorting according to association measures moves NPIs towards the top of the list, as candidates that hardly ever occur in a non-negative context are considered to be highly associated with their negative context. Table 5b illustrates the huge differences between ranking positions of NPIs in the two different lists for three selected NPIs.

5.2 Linguistic Processing

In order to further improve the ordering of the lists, we add more linguistic knowledge to the statistical method. We enriched our result lists with the following linguistically motivated scores:

#NEG	percentage of negative contexts
FIX	degree of morpho-syntactic fixedness
TE	degree of diversity when translated
PDA	percentage of trivial translations

The nature and function of these scores will now be explained one by one. First of all, we use the percentage of the candidates' negative occurrences (#NEG) as a possible indicator for NPIs in our extraction process (cf. Table 6).

	NPI candidate	contexts		freq.	#NEG
+	**aus Kopf gehen**	NEG: 47	NONEG: 0	47	100%
+	**Wald vor Baum sehen**	NEG: 46	NONEG: 4	50	92%
+	**von Fleck kommen**	NEG: 111	NONEG: 14	125	88.8%
-	zu Schaden kommen	NEG: 247	NONEG: 198	445	55.3%

Table 6: Illustration of #NEG score calculation

The morpho-syntactic fixedness score (FIX) is motivated by previous work on the extraction of idiomatic MWEs (Bannard, 2007). Since many multiword NPIs have properties similar to idiomatic expressions, we expect them to be syntactically frozen to a certain degree, which means that they should not permit the usual morphological range of variation of the noun with respect to syntactic features like number, or the use

[14] This is different from the task of retrieving MWEs in general, for which ordering a list of patterns by frequency can already lead to good results. The use of association measures can then further improve initial results.

of all syntactically compatible determiners. Recall that during the extraction of the list of potential candidates, information on the nouns' number and their accompanying determiner is retrieved and stored. For each candidate, we compute the frequency distribution of the number values (SG, PL) and possible determiners (e.g. DEF, INDEF, NONE). The highest percentages of both categories are taken to represent the candidate's preferences. In the case of PNV triples, we also measure the distance of verb, noun and preposition, as idiomatic PNV triples tend to be most often (immediately) adjacent.

The FIX score is calculated for each NPI candidate based on the average of

(i) the #NEG score,

(ii) the percentage of number and article setting, and

(iii) in case of PNV triples: the averaged adjacency-scores.

In order to approximate the semantics of NPI candidates, we use translational entropy (TE) and the proportion of default alignments (PDA). Both scores rely on the assumption that multiword NPIs have a non-compositional semantics, which means that they are to be translated as a whole while compositional combinations of the same syntactic form would exhibit literal translations of their components. To model the translations, we take word equivalences from the EUROPARL corpus (Koehn, 2005). Roughly speaking, the TE score indicates the degree of diversity in a word's translation, while the PDA expresses the percentage of literal (or default) translations. Descriptions of the these two scores can be found in Villada Moirón and Tiedemann (2006).

The linguistic scores are used as follows: We take the top 500 of the lists ordered by *poisson* and re-order these lists according to each of the linguistic scores. In order to measure the quality of the different orderings, we use the uninterpolated average precision (UAP, see Manning and Schütze (1999) for details). Figure 3 shows the results for selected syntactic patterns. Note that for the TE and PDA values, we could only use the EUROPARL corpus (30 million words). As a consequence, some of the NPI candidates cannot be assigned either score (TE or PDA), and are thus skipped in the calculation. The rightmost column contains the resulting UAP value when sorted according to a combination of morpho-syntactic fixedness and translational behavior.[15]

sorted by	poisson	NEG	FIX	TE	PDA	TE+PDA+FIX
NV	0.105	0.069	0.121	0.1	0.124	0.157
ANV	0.233	0.26	0.212	0.174	0.165	0.307
PNV	0.118	0.125	0.145	0.103	0.163	0.2

Figure 3: UAP scores for re-orderings of top 500 (*poisson*)

For the NPI candidates of all three patterns, the orderings according to the linguistic score based on both (monolingual) morpho-syntactic and multilingual features outperform the respective *poisson* orderings. The morpho-syntactic and translational features

[15] The maximal UAP index for a perfectly ordered list would be 1.

are independent and thus benefit from each other when combined. While we achieved our goal to enhance the sorting quality of the candidate lists, the improvement is not great. This may be mainly due to the fact that most NPIs in the lists are relatively low-frequent. The TE and PDA score are not designed for low-frequent data, and computing morpho-syntactic preferences is known to work better for high-frequent data as well.

5.3 Remaining Challenges

There are many expressions that collocate with negation but are not grammatically dependent on it. This is partially due to the nature of newspaper text: For the NPNV-triple *Zugang zu Trinkwasser haben* ('to have access to potable water') we found 29 occurrences all of which appear in a negative context. This is straightforward to explain if we consider that we do not expect a journalist to write about existing access to potable water.

Another obstacle for the statistical approach are contexts that we still cannot model reliably, as well as creative use of language: The NPI *Wald vor Baum sehen* (lit. 'not to see the forest for the trees'; 'not to see the obvious') (contained in Table 5 and in Table 6) occurred 46 of 50 times in a straightforwardly negative context. The complete expression, as it might be listed in a dictionary, is *den Wald vor lauter Bäumen nicht sehen*. In this basic citation form the verb is negated by the sentential negation adverb, *nicht*. Interestingly, this is the form that we observe in 46 cases.

The remaining four occurrences are more difficult: The first, a question (14), is a known nonveridical licensing environment (which we modelled), while the second and third occurrences are a modal context (15) and a conditional clause (16), which are not among the contexts we modeled. In the last sentence, however, there is no clear licensing context at all (17). Regardless of the lack of an obvious negative licensing environment, the sentence is well-formed.[16]

(14) Sieht er dann den Wald vor lauter Bäumen?
see he then the forest despite all trees
'Does he see the obvious?'

(15) Doch wie immer sollte man zunächst einmal den Wald vor Bäumen
but as always should one first the forest despite trees
sehen.
see
'As always one should first note the obvious.'

[16] One might want to speculate that the well-formedness of (17) is contingent on the existence of a presupposition denying that M. L. manages to see the obvious. It has been observed before in the literature that at least some NPIs tolerate this type of indirect and possibly contextual licensing (Richter and Soehn, 2006, pp. 338–339).

(16) Hätte die Kommission eindeutige und anerkannte Prioritäten und könnte
had the commission clear and recognized priorities and could
sie den Wald vor Bäumen sehen, hätten wir nicht diese Aussprache
it the forest despite trees see, had we not this meeting
heute Nachmittag.
today afternoon
'If the commission had clear and recognized priorities and if it could see the obvious, we would not have to meet this afternoon.'

(17) Manchmal sieht M. L. vor lauter Bäumen dennoch den Wald.
sometimes sees M. L. despite all trees still the forest
'Occasionally M. L. does manage to see the obvious.'

It is necessary to keep in mind that for the recognition of each type of negative context, a syntactic pattern of this context—be it a question, some form of conditional or a preceding inherently negative verb—has to be specifically implemented. The examples above illustrate negative contexts that are not easy to detect automatically. As shown in (17), in some cases we might even find constructions with clear NPIs that are used in contexts which cannot be easily categorized as being negative.

6 Results and Discussion

CoDII, the largest collection of German NPIs, comprises 165 entries. Subtracting duplicates that occur in different extraction patterns, our method retrieved 141 NPIs. 25 of these are in CoDII, 116 are new.[17] To appreciate the effectiveness of our method, consider a 'normal-sized' list such as John Lawler's collection of English NPIs[18], which is meant to be an exhaustive listing for English and comprises roughly three dozen entries. Jack Hoeksema's collection of Dutch expressions with strong association to negative environments, which is by far the largest known collection of NPI-like items and has been developed for 15 years, reportedly contains 670 entries.[19] However, Hoeksema's collection is not limited to grammatical NPIs in the narrow sense, i.e. it is not restricted to expressions that are perceived as ungrammatical by native speakers when presented outside of an appropriately negative context. Beyond such expressions, Hoeksema also collects expressions that are statistically strongly associated with negation, which means that they tend not to occur outside of a negative context, although they would still be perceived grammatical if they did.

Lichte and Soehn (2007) do not report how many NPIs their method found. They say that they retrieved 112 items from Kürschner's list of 344 items. However, according to them Kürschner's list contains about 200 pseudo-NPIs, i.e. items which exhibit a

[17] The complete list, including information about which of the retrieved items are in CoDII, is in the appendix.
[18] www-personal.umich.edu/∼jlawler/NPIs.pdf (September 2010)
[19] www.let.rug.nl/∼hoeksema/lexicon_bestanden/v3_document.htm (retrieved in September 2010)

high collocational association with negation but can still occur felicitously in contexts without negation (which makes the empirical scope of Kürschner's list comparable to Hoeksema's). Given that Lichte and Soehn's extraction algorithm primarily targeted single-word NPIs, and that all NPIs that they identified are in CoDII, the overlap between the items they extracted and ours must be small (equal to or below 25).

The main reason for the small overlap should come from the different strategies of selecting multiword NPI candidates. Our procedure is based on syntactically meaningful patterns which enter statistical processing as a whole. Lichte's extraction procedure starts with single lemmata which are extended one lemma at a time, and only those chains of length $n + 1$ that exhibit higher association to negation than the nucleus chain of length n are further considered. Apart from the computational inefficiency of considering arbitrary other lemmata (in the same clause) for extending a lemma chain, this procedure can only suggest as candidates those expressions of length m such that a sequence of of chains exists where each succeeding chain has a higher association to negation than its shorter predecessor. The existence of such a sequence of chains cannot be guaranteed for each multiword NPI, and if it does not exist, Lichte's procedure will not find the NPI.

The types of NPIs found with the three most successful search patterns, NV (29), PNV (31), and ANV (77) show interesting differences. The PNV list contains a high number of idiomatic expressions (*(mit etw.) hinter dem Berg halten, (sich) in die Karten schauen (lassen), auf den Mund gefallen (sein)*), and only a small number of semantically transparent MWEs (*mit Vorwürfen sparen, mit keinem Wort erwähnen*). The list NV is similar, containing a somewhat smaller but still sizable number of non-decomposable idioms. Finally, the third list, ANV, is markedly different, containing mostly non-idiomatic, semantically transparent MWEs such as *wesentliche Änderungen erwarten, (sich) einen anderen Rat wissen, eine andere Wahl sehen*, and a smaller number of clearly idiomatic expressions (*einen blassen Schimmer haben, schlafende Hunde wecken*). These differences between the lists could explain the varying success with reordering the top 500 by taking linguistic knowledge about the fixedness of expressions into account. The more idiomatic an expression is, the more restricted is its syntactic flexibility, and our linguistic processing favors syntactically frozen expressions. The NPIs found in the PNV pattern should thus be promoted more than the ones in the ANV pattern. A last, more general observation concerns the tendency in the ANV pattern (also visible in the PNV pattern) of expressions forming clusters which only differ in the verb, such as *geringsten Zweifel {aufkommen | geben | haben | hegen | lassen}*. From a theoretical point of view, one might be tempted to consider these items variants of one and the same underlying NPI, and it might be interesting to investigate the types of verbs that can enter into these variants, and their properties.

Our success quota of finding true NPIs in five top 500 lists (141 in 2500) clearly shows that we have not designed a fully automatic NPI extraction procedure. There is considerable human effort and expert judgment involved in finding NPIs in candidate lists. This state of affairs echoes the early apprehensions by Hoeksema (1997), who feared that difficulties such as those arising from inaccurate recognition of licensing

environments or from polysemy of words and ambiguity of constructions could defeat automatizing NPI detection. Looking at the experience gathered with the two candidate extraction procedures that have been designed in the meantime, we agree that what we have achieved is probably very coarse-grained and might even suffer from inherent shortcomings that cannot be completely overcome by simply pursuing the same strategy further. Despite these imperfections, the methods that we applied are still highly successful insofar as they contribute dramatically to improving our database of German NPIs. As this is still only a beginning, we can hope for many more NPIs to be found by considering other promising syntactic patterns and by refining the candidate extraction procedure. In particular, we only applied the linguistic processing methods of Section 5.2 to candidate lists after they were annotated by human experts. Since linguistic processing improved the rankings in these pre-sorted candidate lists, it should be checked if their application at an earlier stage in the extraction pipeline improves the candidate lists given to human annotators.

7 Conclusion and Future Work

As we mentioned several times, many NPI licensing environments exhibit the logical property of being downward entailing, which means that they support inferences from supersets to subsets (see the example in Section 2). For this reason, detecting downward entailing environments is highly relevant for determining textual entailments. In a recent paper, Danescu-Niculescu-Mizil et al. (2009) exploit the licensing requirements of NPIs and use a set of English NPIs to extract downward-entailing operators from text. In a sense, this is the converse task to ours, but it presupposes a lexicon of NPIs. Knowledge of a larger set of NPIs in a given language, as provided by our method, should help improve extraction of downward-entailing operators, and may thus ultimately contribute to improving textual entailment tasks in language processing.

We showed that by sorting candidate-context pairs according to their log-likelihood scores, NPIs could be retrieved with considerable precision. In a second step, we applied linguistically motivated scores in order to enhance sorting quality for the top 500 entries of the log-likelihood sorting. We saw that our results were very promising, as we managed to increase the number of known NPIs in German by more than two thirds. However, we also believe that there is still much room for improvement by integrating linguistic knowledge and statistical processing more tightly. With a more fine-grained definition of negative contexts, as provided by the linguistic literature, we would hope to obtain better candidate lists.

Looking at our results from the perspective of theoretical linguistics, there should be much to gain from semi-automatic NPI extraction methods. Many questions about the syntactic, semantic and pragmatic nature of NPIs and their licensing environments are still open. Having a much larger empirical base for investigating these issues should contribute significantly to improving the linguistic theory. For example, one major hypothesis of pragmatic NPI theories claims that their behavior can be attributed to their property of being minimizers. This seems hard to maintain considering that many

items of the joined item list consisting of CoDII and our newly extracted items are not end-of-scale elements in any obvious way. Expressions such as *jemandem (nicht) grün sein* (lit. 'so. (not) green be'; '(not) to be well-disposed toward someone'), *jemandem (nicht) über den Weg trauen* (lit. 'so. (not) over the path trust'; '(not) to trust someone') or *jemandem (nicht) von der Seite weichen* (lit. 'so. (not) from the side leave'; 'to tag along after someone') are not at an endpoint of any easily imaginable scale; many similar examples can be found by simply going through the list. Any claim to universality of theories that explain NPIs from their supposed property of being minimizers seems to be doomed considering the full range of data.

Having a large repository of NPIs also opens up new avenues for psycholinguistic research. Among the problems of researching properties of NPIs such as the distinction between *weak* and *strong* NPIs (i.e. between those items that are satisfied with weaker forms of negation in their licensing contexts and those which require stronger forms of negation) has been the diversity of syntactic categories and syntactic form of multiword NPIs. In psycholinguistic experiments we typically want to vary exactly one feature under investigation to make sure that other variation does not interfere or mask those effects that we want to study. With only a dozen or two NPIs, it is very hard or impossible to construct enough items for an experiment which only vary in one dimension. The kind of NPI database that we have now compiled makes it much easier to address the kinds of questions that psycholinguists and linguists might want to ask about the nature of NPIs, because we are now in a much better position to construct syntactically more uniform item sets for experiments. Richter and Radó (2010) have already used items from our extraction procedure in the study of demonstrating the psycholinguistic reality of the weak/strong classification, the behavior of strong and weak NPIs in Neg-Raising contexts (Horn, 1978), and so-called intervention effects of proportional quantifiers in Neg-Raising constructions with NPI licensing. These experiments would not have been possible without a large resource of syntactically similar NPIs. For this very reason, corresponding experiments can currently not be conducted for English.

Our success with finding many previously unobserved NPIs among five patterns of MWEs supports our initial intuition of a deeper relationship between the property of being an NPI and idiomatic expressions. In normal idiomatic expressions there is a strong association between the set of words that make up the idiomatic expressions, whereas multiword NPIs additionally exhibit a strong association to a more abstract grammatical feature, *viz.* (various degrees of) negation. Investigating the relationship between these apparently distinct types of grammatical association might reveal interesting, hitherto unnoticed properties of idiomatic expressions, and might lead to a re-evaluation of the function of NPIs in the grammatical system.[20]

[20] The theoretical implications of these considerations are pursued further in Richter et al. (2010).

Acknowledgments

We are very grateful to Manfred Sailer for his generous help in checking our lists of NPI candidates for true NPIs, and to Timm Lichte for giving us very useful background information on his work on NPI extraction, including the original Perl script of the NPI extraction algorithm of Lichte and Soehn (2007) with his complete definition of licensing environments. Thank you to Janina Radó for proofreading.

References

Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the ACL Workshop on a broader perspective on multiword expressions*, pages 1–8, Prague, Czech Republic.

Chierchia, G. (2006). Broaden your views. Implicatures of domain widening and the 'logicality' of language. *Linguistic Inquiry*, 37(4):535–590.

Danescu-Niculescu-Mizil, C., Lee, L., and Ducott, R. (2009). Without a 'doubt'? Unsupervised discovery of downward-entailing operators. In *Proceedings of NAACL HLT*, pages 137–145.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Fauconnier, G. (1975). Pragmatic scales and logical structure. *Linguistic Inquiry*, 6(3):353–375.

Fritzinger, F. (2009). Using parallel text for the extraction of German multiword expressions. *Lexis - E-journal in English Lexicology*, 4.

Fritzinger, F. and Heid, U. (2009). Automatic grouping of morphologically related collocations. In *Online Proceedings of the Corpus Linguistics Conference 2009*, Liverpool/UK.

Fritzinger, F., Richter, F., and Weller, M. (2010). Pattern-based extraction of negative polarity items from dependency-parsed text. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*. European Language Resources Association.

Hoeksema, J. (1997). Corpus study of negative polarity items. Html version of a paper which appeared in the *IV-V Jornades de corpus linguistics 1996–1997*, Universitat Pompeu Fabre, Barcelona. URL: www.let.rug.nl/hoeksema/docs/barcelona.html.

Horn, L. R. (1978). Remarks on Neg-Raising. In Cole, P., editor, *Pragmatics*, volume 9 of *Syntax and Semantics*, pages 129–220. Academic Press, New York, San Francisco, London.

Kadmon, N. and Landman, F. (1993). Any. *Linguistics and Philosophy*, 16:353–422.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th MT Summit 2005*, Phuket, Thailand.

Krifka, M. (1995). The semantics and pragmatics of polarity items. *Linguistic Analysis*, 25:209–257.

Kürschner, W. (1983). *Studien zur Negation im Deutschen*. Gunter Narr, Tübingen.

Ladusaw, W. A. (1980). On the notion 'affective' in the analysis of negative-polarity items. *Journal of Linguistic Research*, 1(2):1–16.

Lichte, T. (2005a). Corpus-based acquisition of complex negative polarity items. In Gervain, J., editor, *Proceedings of the Tenth ESSLLI Student Session*, Edinburgh. Heriot-Watt University.

Lichte, T. (2005b). Korpusbasierte Acquirierung negativ-polärer Elemente. Master's thesis, Seminar für Sprachwissenschaft, University of Tübingen.

Lichte, T. and Soehn, J.-P. (2007). The retrieval and classification of negative polarity items using statistical profiles. In Featherston, S. and Sternefeld, W., editors, *Roots: Linguistics in Search of its Evidential Base*, pages 249–266. Mouton de Gruyter, Berlin.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Richter, F. and Radó, J. (2010). NPI licensing in German: Some experimental results. Unpublished manuscript. Eberhard Karls Universität Tübingen. October 2010.

Richter, F., Sailer, M., and Trawiński, B. (2010). The collection of distributionally idiosyncratic items: An interface between data and theory. In Ptashnyk, S., Hallsteinsdóttir, E., and Bubenhofer, N., editors, *Korpora, Web und Datenbanken. Computergestützte Methoden in der modernen Phraseologie und Lexikographie*, volume 25 of *Phraseologie und Parömiologie*, pages 247–261. Schneider Verlag Hohengehren GmbH.

Richter, F. and Soehn, J.-P. (2006). 'Braucht niemanden zu scheren': A survey of NPI licensing in German. In Müller, S., editor, *Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*, pages 421–440. CSLI Publications.

Schiehlen, M. (2003). A cascaded finite-state parser for German. In *Proceedings of the 10th EACL*, Budapest, Hungary.

Trawiński, B., Soehn, J.-P., Sailer, M., and Richter, F. (2008). A multilingual electronic database of distributionally idiosyncratic items. In Bernal, E. and DeCesaris, J., editors, *Proceedings of the XIII Euralex International Congress*, volume 20 of *Activitats*, pages 1445–1451, Barcelona, Spain. Universitat Pompeu Fabra.

van der Wouden, T. (1997). *Negative Contexts. Collocation, Polarity and Multiple Negation*. Routledge, London.

Villada Moirón, B. and Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word alignment. In *Proceedings of the EACL 2006 Workshop on multiword-expressions in a multilingual context*, Trento, Italy.

Weller, M. and Heid, U. (2010). Extraction of German multi-word expressions from parsed corpora using context features. In *Proceedings of the Linguistic Resources and Evaluation Conference, LREC 2010*, Valetta, Malta.

Zwarts, F. (1995). Nonveridical contexts. *Linguistic Analysis*, 25:286–312.

Zwarts, F. (1997). Three types of polarity. In Hamm, F. and Hinrichs, E. W., editors, *Plurality and Quantification*, pages 177–237. Kluwer Academic Publishers, Dordrecht.

Appendix

This appendix lists all NPIs that were extracted from the corpora in Table 1 and confirmed by human annotators of the candidate lists. They are sorted according to our five syntactic patterns. For each NPI, the tables show if it is already contained in the CoDII collection ('+'), or if it is a new NPI not noted there ('–'). The annotation '(+)' marks partial NPIs, i.e. expressions that are not complete NPIs yet but can be recognized as parts of true NPIs contained in CoDII. For example, the PNV pattern contains *über Tatsache hinwegtäuschen*, which expands to the NPI *über Tatsache hinwegtäuschen können*. With adjectives, 'C' denotes comparative morphology, and 'S' denotes superlative forms.

PNV	in CODII
vor Anfrage retten	–
vor Auftrag retten	–
hinter Berg halten	–
mit Ding zugehen	–
von Eltern sein	(+)
von Fleck kommen	–
in Haut stecken	–
in Karte schauen	(+)
aus Kopf gehen	–
mit Kritik sparen	–
in Moment sagen	–
auf Mund fallen	(+)
vor Mund nehmen	(+)
hinter Ofen hervorlocken	–
an Schlaf denken	–
von Seite weichen	+
mit Silbe erwähnen	–
aus Staunen herauskommen	+
von Stelle kommen	–
auf Stuhl halten	–
über Tatsache hinwegtäuschen	(+)
in Traum denken	+
in Traum einfallen	–
ohne Tücke sein	–
mit Vorwurf sparen	–
über Weg trauen	–
in Weise entsprechen	–
in Weise rechen	–
an Wiege singen	–
mit Wimper zucken	+
mit Wort erwähnen	–

NV	in CODII
Abbruch tun	+
Ahnung haben	+
Aufschub dulden	–
Berührungsangst kennen	–
Blumentopf gewinnen	+
Erbarmen kennen	–
Finger rühren	+
Haar krümmen	+
Haar lassen	(+)
Halten geben	–
Hauch haben	–
Hehl machen	+
Kosten scheuen	(+)
Mördergrube machen	(+)
Mühe scheuen	(+)
Pfennig erhalten	–
Pfennig haben	–
Pfennig sehen	–
Pfennig zahlen	–
Pfifferling geben	–
Rede sein	(+)
Sekunde zweifeln	–
Stein lassen	(+)
Tabu kennen	–
Träne nachweinen	–
Welt verstehen	(+)
Wort glauben	–
Wort verlieren	–
Wort verstehen	–

ANV	CODII
geringS Abstrich machen	–
geringS Ahnung haben	+
leisS Ahnung haben	+
geringS Anhaltspunkt geben	–
geringS Anlass geben	–
ganz Aufregung verstehen	–
weitC Aufschub dulden	–
ander Ausweg lassen	–
ander Ausweg sehen	–
ander Ausweg wissen	–
nennenW Auswirkung haben	–
recht Bezug finden	–
ander Chance haben	–
ander Chance sehen	–
geringS Chance haben	–
gewiß Charme absprechen	–
blass Dunst haben	–
gut Faden lassen	–
geringS Einfluss haben	–
nennenW Einfluss haben	–
recht Freude haben	–
groß Gedanke machen	–
geringS Grund sehen	–
einzig Grund geben	–
einzig Grund haben	–
erkennenB Grund geben	–
geringS Grund geben	–
zwingend Grund geben	–
gut Haar lassen	+
schlafende Hund wecken	–
geringS Hinweis finden	–
geringS Hinweis geben	–
groß Illusion machen	–
geringS Interesse haben	–
sonderlich Interesse haben	–
gut Licht werfen	–
geringS Lust haben	–
recht Lust haben	–
ander Möglichkeit sehen	–

ANV	CODII
geringS Problem haben	–
ander Rat wissen	–
recht Reim machen	–
gering Rolle spielen	–
nennenW Rolle spielen	–
gutS Ruf haben	–
gutS Ruf genießen	–
halb Sache machen	–
blass Schimmer haben	+
ganz Schritt halten	–
klein Seitenhieb verkneifen	–
recht Sinn ergeben	–
recht Spaß machen	–
groß Sprung erlauben	–
groß Sprung machen	–
groß Sprung zulassen	–
leicht Stand haben	–
gutS Tag erwischen	–
nennenW Unterschied geben	–
nennenW Veränderung erwarten	–
geringS Verständnis haben	–
ander Wahl bleiben	+
ander Wahl geben	–
ander Wahl haben	–
ander Wahl lassen	–
ander Wahl sehen	–
gut Wille absprechen	–
eigen Wort verstehen	–
einzig Wort verlieren	–
einzig Wort verstehen	–
groß Wort verlieren	–
weitC Wort verlieren	–
geringS Zweifel aufkommen	–
geringS Zweifel geben	–
geringS Zweifel haben	–
geringS Zweifel hegen	–
geringS Zweifel lassen	–
leisS Zweifel hegen	–

NPNV	in CODII
Anspruch auf Vollständigkeit erheben	–
Blatt vor Mund nehmen	+
Mördergrube aus Herz machen	+
Wald vor Baum sehen	–
Zweifel an Haltung lassen	–

PANV	in CODII
über mangelnde Arbeit beklagen	–
von schlecht Eltern sein	+
in kühnS Traum erwarten	–
auf grün Zweig kommen	+

For the lists of items above, it is important to note that many of the newly found items are partial NPIs: For example, in the PNV pattern, the verb *können* has to be appended to each of the first two items (*vor Anfrage retten, vor Auftrag retten*) and *mit Ding zugehen* has to be extended to *mit recht Ding zugehen* in order to obtain complete NPIs; among the first 4 items on this list only *hinter Berg halten* is already complete in the form in which it was extracted with the PNV pattern.

Most NPIs should be relatively easy to recognize from the form in which they occur on the lists. For *in Weise rechen* and *an Wiege singen* it might be somewhat harder to identify their citation forms, *in (k)einer Weise gerecht werden* and *an der Wiege gesungen sein*.

Examples of NPIs that occur partially in one list and completed in another are *von (schlecht) Eltern sein* (PNV, PANV), *(Blatt) vor Mund nehmen* (PNV, NPNV), *(gut) Haar lassen* (NV, ANV), and *Mördergrube (aus Herz) machen* (NV, NPNV). All four appear in CoDII. A special case are two partial items in the NV pattern, *Kosten scheuen* and *Mühen scheuen*, which combine to the complete NPI *Kosten und Mühen scheuen*, which belongs to a pattern that we were not investigating here.

The items *Ahnung haben, Aufschub dulden, Wort verlieren* and *Wort verstehen* in the NV pattern occur in extended forms in the ANV pattern, three of them with several extensions. We consider their NV forms independent complete NPIs, as they are found in corpora in these short forms. However, the longer forms also seem to be independent collocational units, which justifies listing them separately in larger patterns.

Author Index

Steffen Eger
IDS Mannheim
eger.steffen@googlemail.com

Bryan Jurish
Berlin-Brandenburgische Akademie der Wissenschaften
jurish@ling.uni-potsdam.de

Tim vor der Brück
Fernuniversität Hagen
Tim.vorderBrueck@fernuni-hagen.de

Hermann Helbig
Fernuniversität Hagen
hermann.helbig@fernuni-hagen.de

Yvonne Zajontz
DHBW Stuttgart
zajontz@dhbw-stuttgart.de

Marc Kuhn
DHBW Stuttgart
marc.kuhn@dhbw-stuttgart.de

Vanessa Kollmann
DHBW Stuttgart
kollmann@dhbw-stuttgart.de

Frank Richter
Seminar für Sprachwissenschaft, Universität Tübingen
fr@sfs.uni-tuebingen.de

Fabienne Fritzinger
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
fritzife@ims.uni-stuttgart.de

Marion Weller
IInstitut für Maschinelle Sprachverarbeitung, Universität Stuttgart
wellermn@ims.uni-stuttgart.de