An XML-based Tool for Tracking English Inclusions in German Text

Beatrice Alex[1] and Claire Grover[1]

University of Edinburgh, School of Informatics, 2 Buccleuch Place,
Edinburgh, EH8 9LW
{v1balex,grover}@inf.ed.ac.uk

Abstract. The use of lexicons and corpora advances both linguistic research and performance of current natural language processing (NLP) systems. We present a tool that exploits such resources, specifically English and German lexical databases and the World Wide Web to recognise English inclusions in German newspaper articles. The output of the tool can assist lexical resource developers in monitoring changing patterns of English inclusion usage. The corpus used for the classification covers three different domains. We report the classification results and illustrate their value to linguistic and NLP research.

1 Introduction

The increasing influence which English is having on German, sometimes referred to as *Denglish* (German mixed with English), has developed into a controversial topic widely discussed in the German media and even appeared on Germany's political agenda [10]. Whether accepted by native speakers or not, this phenomenon is studied by linguists and lexicographers for whom an automatic classifier of foreign inclusions would prove a valuable tool. It must also be dealt with by developers of NLP systems as the correct handling of foreign inclusions would be beneficial to machine translation (MT) and text-to-speech (TTS) processing.

Existing language identification systems are not very accurate for individual words (e.g. [7]). We have developed a robust and highly efficient tool that identifies English inclusions in German text on the word level by means of a computationally inexpensive lookup procedure. This system allows linguists and lexicographers to observe language in use, including changes over time, and to investigate the use and frequency of loan words in a given language and domain.

In Section 2 we address the background for this research in detail. Section 3 describes the corpus used for the experiment and provides a system overview. In Section 4 we report our results and analyse various sources of error. Finally, we present our conclusions and outline directions for future work in Section 5.

2 Background

This paper describes a preliminary research initiative to study the types of foreign inclusions, in particular English inclusions in German text. As English is

currently the dominant language of business, science & technology, advertising and other sectors, it has become one of the main sources of borrowing. The first anglicisms appeared in German during the Middle Ages [17]. However, English has primarily influenced German during the 19th and 20th centuries. In the second half of the 20th century, an enormous increase in the number of anglicisms entering German was recorded. This can be mainly attributed to political events such as the creation and enlargement of the EU as well as technological advances, in particular the invention of the computer and internet. As a result, German documents frequently contain English names and expressions. Lexical resources need to be updated to reflect this trend.

Our system was built to examine the frequency of English inclusions in German newspaper text on different subjects and to gain a better understanding of how to recognise such instances automatically. Foreign inclusions can be regarded as borrowings which are further sub-divided into (1) assimilated loan words which are relatively integrated into the receiver language and (2) foreign words which are integrated into the receiver language to a lesser extent [20].

Our system is specifically tailored to recognise *foreign words* stemming from English. However, the system also identifies words with the same spelling in both languages, including many assimilated loan words and internationalisms stemming from English and other languages. Borrowings also include loan substitutions (Lehnprägungen [2]) or internal borrowing (inneres Lehngut [20]) such as *Spracherkennung* (speech recognition). These are instances where the lexical items of the donor language are expressed using semantically identical or similar lexical items of the receiver language. Loan substitutions are not separately identified and, for the purpose of our experiment, classified as German words.

Our processing paradigm is XML-based. As a markup language for NLP tasks, XML is expressive and flexible yet constrainable. Furthermore, there exist a wide range of XML-based tools for NLP applications which lend themselves to a modular, pipelined approach to processing whereby linguistic knowledge is computed and added incrementally as XML annotations. Moreover, XML's character encoding capabilities facilitate multilingual processing. Our tool for processing German text is essentially a pipeline which converts HTML files to XML and applies a sequence of modules to add linguistic markup and to classify nouns as German or English. The pipeline is composed of calls to a variety of XML-based tools from the TTT and LTXML toolsets [9, 16]. In addition, we have integrated non-XML public-domain tools (e.g. the TnT tagger, [3]) and incorporated their output into the XML markup. The XML output can be searched to find specific instances or to acquire counts of occurrences.

3 Corpus and System Description

For our experiment, we used a selection of German newspaper articles and a set of different modules combined in a UNIX pipeline. These modules are: a text pre-processing system, including tokenisation and part-of-speech (POS) tagging, a lexicon lookup and a Google lookup. The main advantage of this setup is the

ability to integrate the output of new modules specifically tailored to our task with that of already existing tools in an organised fashion.

3.1 Corpus and Domains

The text corpus is made up of a random selection of newspaper articles published in the Frankfurter Allgemeine Zeitung[1] between 2001 and 2004 in the domains of (1) *internet & telecomms*, (2) *EU* and (3) *space travel*. With approximately 16,000 tokens per domain, the overall corpus comprises of 48,000 tokens. The specific domains were chosen to examine the use and frequency of English inclusions in German texts of a more technological, political or scientific nature.

3.2 Pre-processing

The downloaded webpages are firstly processed using `Tidy`[2] to remove HTML markup and then converted into XML. The resulting XML pages simply contain the textual information of each article. The corpus is subsequently tokenised by means of two rule-based grammars which we developed specifically for German. The first grammar pre-tokenises the text into tokens surrounded by white space and punctuation and the second grammar resolves various abbreviations, numerals and URLs. Both grammars function in conjunction with `Lxtransduce`[3], a transducer which processes an input stream and rewrites it based on the rules provided by adding or rewriting XML markup. `Lxtransduce` is a recently updated version of `Fsgmatch`, the core program of the Language Technology Text Tokenisation Toolkit [9]. The tokenised output is POS-tagged using TnT.

3.3 Lexicon Lookup

We used CELEX[4], a lexical database of English, German and Dutch, for the initial lookup. The English database contains 52,446 lemmas representing 160,594 corresponding word forms and the German database holds 51,728 lemmas and their 365,530 word forms. CELEX lookup was carried out only for tokens that TnT tagged as nouns and foreign material. Several studies [20, 21, 4] have shown that of all anglicisms found in German texts, nouns were the most frequent ones, accounting for more than 90% of tokens in the domains of general news and computing. Anglicisms representing other parts of speech are relatively infrequently used. This is the reason for focussing on the classification of nouns, although one future objective is to extend the lookup to other parts of speech as well.

Each token was looked up twice, both in the German and English CELEX databases, using `grep`. Each part of hyphenated compounds was checked individually. Moreover, we made the lookup in the English database case-insensitive in order to identify the capitalised English tokens in our corpus, the reason being

[1] `http://www.faz.net`
[2] http://tidy.sourceforge.net
[3] `http://www.ltg.ed.ac.uk/~richard/lxtransduce.html`
[4] `http://www.kun.nl/celex/`

Table 1. Most frequent words per domain found in both lexicons

Internet & Telecoms		**European Union**		**Space Travel**	
Token	Frequency	Token	Frequency	Token	Frequency
Dollar	16	Union	28	Station	58
Computer	14	April	12	All	30
Generation	12	Referendum	10	Start	27
April	12	Fall	9	Mission	16
Autos	7	Rat	8	Chef	14

that all proper and regular nouns are capitalised in German. The lexicon lookup is also sensitive to POS tags to reduce classification errors.

On the basis of this initial lookup, each token was either found only in the German lexicon, only in the English lexicon, in both or in neither lexicon. The majority of tokens found exclusively in the German lexicon are either German words or English words with German case inflections such as *Computern*. The word *Computer* is now used so frequently in German that it has already been entered into lexicons and dictionaries. In order to detect the base language of the latter, a second lookup was performed to check whether the lemma of the token (given in CELEX) is also found in the English lexicon. For example, in the case of *Computern*, the lemma *Computer* was found. Tokens found exclusively in the English lexicon such as *Software* or *News* are generally English words and do not overlap with German lexicon entries. A large majority of them are clear instances of foreign inclusions.

Tokens contained in both lexicons include words with the same spelling in both languages (Table 1). These are words without inflectional endings or ending in *s* coinciding with the German genitive singular or the German and English plural forms of that token, e.g. *Computers*. The majority of these lexical items have the same semantics in both languages. A subgroup are evidently English loan words (e.g. *Computer*). Others represent assimilated loans and cognates with the same orthographic characteristics in both languages where the language origin is not always immediately apparent (e.g. *Mission*). This phenomenon is due to the fact that German and English belong to the same language group (Germanic), and have been influenced similarly by other foreign languages including Latin and French[19]. It should also be mentioned that English text contains some German loan words, though to a much lesser extent. Our German corpus contains such an example, the word *Ersatz*, which was found in the English lexicon. Additionally, Table 1 shows that a simple lexicon lookup also detects inter-linguistic homographs with different semantics in either language, including *Fall* (*case* vs. *fall*) and *Rat* (*council/advice* vs. *rat*). A deeper semantic analysis is necessary to distinguish such homographs. All tokens found in neither lexicon are submitted to the Google search engine. These include:

– German compounds, including loan substitutions: *Mausklick* (mouse click)

– English unhyphenated compounds: *Homepage*, *Hypertext*, *Spacehab*
– Mixed-lingual unhyphenated compounds: *Shuttleflug* (shuttle flight)
– English nouns with German inflections: *Receivern*
– Abbreviations and acronyms: *UMTS*, *UKW*
– Words with spelling mistakes: *Abruch* (abortion)
– English words with American spelling: *Center*

3.4 Google Lookup

This lookup exploits the fact that the Web is a large resource with textual material in a multiplicity of languages. Being a US-innovation, it was originally a completely English medium. A study carried out by the Babel project[5] showed that in 1997 82.3% of a set of 3239 selected webpages were written in English, 4.0% in German, followed by small percentages of webpages in other languages. Since then, the estimated number of webpages written in languages other than English has increased, which means that the Web presence of languages is becoming more and more reflective of their distribution in the real world [5, 6].

A novel trend in computational linguistics has been the utilisation of the Web as a linguistic corpus. Although the information published on the Web is sometimes noisy, its sheer size and the continuous addition of new material make it a valuable pool of information in terms of languages in use. The Web has already been used successfully for several NLP tasks such as NE acquisition [11], disambiguation of prepositional phrase attachments [18], anaphora resolution [13], word sense disambiguation [1] and MT [8, 15]. For a detailed overview of these experiments see also [12].

The principle motivation for choosing Google for our work is the fact that it searches more than 4 million webpages, a large portion of all information available on the Web. The Google lookup was carried out only for the sub-group of tokens found in neither lexicon in order to keep the computational cost to a minimum. In addition, several smaller experiments showed that the lexicon lookup was already sufficiently accurate for tokens contained exclusively in the German or English databases. Besides, current Google search options are limited in that it is impossible to treat queries case- or POS-sensitively. Therefore, tokens found in both lexical databases would often be wrongly classified as English, particularly those that are frequently used (e.g. *All*).

Therefore, only tokens found neither in the German nor in the English lexicon were submitted to Google. We obtained the number of hits for two searches per token, one exclusively on German webpages and one on English ones, an advanced language preference offered by Google. Each token was classified as being either German or English depending on which search returned more hits. The underlying assumption here is that a German word is more frequently used in German text than in English and, similarly, the use of an English word is more common in English documents. If both searches returned zero hits, the token

[5] `http://www.isoc.org:8030/palmares.en.html`

was classified as German by default. This happened only for two tokens: *Orientierungsmotoren* (navigation engines) and *Reserveammoniak* (spare ammonia).

The following is an example sentence of the system output retaining the language classification alone (DE=German, EN=English):

> Mit diesem <DE>Verhalten</DE> übertragen die stets gut informierten <EN>Smartshopper</EN> den harten <DE>Preiswettbewerb</DE> im <EN>Internet</EN> in die stationären <DE>Geschäfte</DE>.
> *With this behaviour, ever well-informed Smart Shoppers are transferring the strong price competition in the Internet to stationary businesses.*

4 Evaluation and Analysis

4.1 Results

Table 2 presents the number of tokens and types classified as German or English per domain, on the basis of the lookup in Google and/or in the lexicons, as well as the number of tokens and types with the same spelling in both languages. The results show that the combination of the lexicon and the Google lookup is advantageous for a large portion of tokens. This supports our hypothesis that the Web offers valuable linguistic knowledge. There are considerably more English tokens present in the articles on the internet & telecoms (632 = 17%) and space travel (340 = 9%) than in those on the EU (94 = 3%). This result seemed surprising at first as the development of the EU has facilitated increasing contact between German and English speaking cultures. However, political structures and concepts are intrinsic parts of individual cultures and therefore tend to have their own expressions. Moreover, EU legislation is translated into all its

Table 2. Token and type statistics per domain

Domain	**Internet & Telecoms**		**European Union**		**Space Travel**	
	Tokens	**Types**	**Tokens**	**Types**	**Tokens**	**Types**
Total	**15919**	**4386**	**16028**	**4200**	**16066**	**4126**
Looked up	**3780**	**1735**	**3371**	**1615**	**3680**	**1522**
Lexicon	2371	980	2479	1056	2722	1019
Lex. + Google	1409	755	892	559	958	503
Classif. as DE	**2922**	**1461**	**3077**	**1493**	**2961**	**1353**
Lexicon	2016	822	2264	966	2247	894
Lex. + Google	906	639	813	527	714	459
Classif. as EN	**632**	**180**	**94**	**45**	**340**	**71**
Lexicon	129	64	15	13	96	27
Lex. + Google	503	116	79	32	244	44
Same spelling	**226**	**94**	**200**	**77**	**379**	**98**

Table 3. Most frequent English words per domain

Internet & Telecoms			**European Union**			**Space Travel**		
Token	Freq.	GOOGLE	Token	Freq.	GOOGLE	Token	Freq.	GOOGLE
Internet	106	✓	DCEI	10	✓	Shuttle	32	
Online	64	✓	Cluster	3		Crew	19	
UMTS	24	✓	Spreads	1		US	14	✓
Handy	13	✓	Scores	1		Shuttles	7	
PC	12	✓	Portfolio	1		Space	2	

official languages, which have risen from 11 to 20 since the recent enlargement. This language policy indicates that English is less dominant in this domain than expected. The strong presence of English inclusions in the articles from the other two domains was anticipated, as English is the dominant language in science & technology.

Table 3 lists the most frequent English terms identified by our system. As we aim to illustrate how our tool aids the discovery of emerging anglicisms in German, we removed tokens from the list that were mistakenly classified as English due to wrong POS tagging and other errors. Such cases are discussed separately in the error analysis (Section 4.2). Table 3 includes various types of anglicisms some of which were identified by the lexicon lookup alone and others for which the Google lookup proved beneficial. Firstly, there are English terms whose German equivalents are rarely used, such as *Internet* (Netz). This is reflected in their low frequency in our corpus, e.g. *Netz* only appeared 25 times. This result corresponds to the findings by Corr [4] which show that Germans tend to favour the use of anglicisms referring to specific computer vocabulary to that of their German translations. Table 3 also contains examples of English words that have well established and frequently used German equivalents such as *Shuttle* (Raumfähre). The German translation of this example occurred 47 times. In this case, the German word was used 59% and the English equivalent 41% of the time. Our tool can be re-run on new data to examine whether the percentages of English inclusions increase over time. A further interesting example listed in Table 3 is *Handy*, the word used by Germans for *mobile phone*. It was classified as English according to the counts of the Google query which is insensitive to case and POS tags. This word is a pseudo loan, a type of borrowing that is pronounced as the lexical item of the donor language but where the meanings in the donor and receiving languages differ. Although linguists disagree on whether pseudo loans can be classed as borrowings, in this case an anglicism, it is clear that such instances would not exist in the receiving language if they had not been derived from the lexical item in the donor language. The word *Handy* originated from the *Handy Talkie*, the first hand-held two-way radio developed in 1940 [14].

English abbreviations such as *PC* - personal computer (Table 3) represent specific cases of assimilated loan words that are phonologically integrated in German. Our system classified such examples as English as they occur more

frequently on English webpages. Other abbreviations identified as English are *HTML* - Hypertext Markup Language, *WWW* - World Wide Web and *GPS* - Global Positioning System. Similarly, abbreviations for German expansions were classified as German, e.g. *UKW* - Ultrakurzwellen (frequency modulation), *EZB* - Europäische Zentralbank (European Central Bank) and *MEZ* - mitteleuropäische Zeit (Central European Time). As this classification failed rarely (Section 4.2), it represents a relatively reliable indicator of whether abbreviations expand to English or German terms.

All aforementioned examples demonstrate the increasing influence that English has on German. Our system can be easily tested on a new corpus at a later stage to monitor this influence over time. Such information is particularly valuable for linguistic and lexicographical studies on the latest language changes and the emergence of new vocabulary.

4.2 Error Analysis

Although a careful examination showed that the system produces relatively accurate output, there are a number of instances where errors occur (Table 4). The majority of errors are caused by the wrong assignment of POS tags, in particular for NEs referring to organisations, persons and locations that were tagged as common nouns. Moreover, the system is not designed to determine the individual English and German morphemes of mixed-lingual unhyphenated compounds directly. When searched in Google, they always receive a higher number of hits for German webpages. If the English morpheme of such a compound occurs in the corpus as a separate lexical item, a subsequent search allows us to find all the mixed-lingual compounds which it is part of. For example, a search of the word *Shuttle* in the group of words classified as German returns *Shuttleflug* and *Shuttlestart*. This additional lookup is, however, not an option if the English morpheme does not occur individually, as is the case for *Vertragsjobs*. A deeper morphological analysis is required to tackle this problem. Other less frequent language classification errors are caused for new internationalisms that have not yet been entered into the lexical databases and for abbreviations with several expansions in different languages. An example for the latter is *BIP* which stands for *Bruttoinlandsprodukt* (gross domestic product) but which also has several expansions in English including *business investment planning* or *business incentive policy*. One clear example of a misleading Google count occurred for *Kameratelefon*, where more hits were returned for English webpages, but such errors were rare. In future, the number of Google hits could be applied as a weight in a machine learning classification system. Finally, our tool unsurprisingly failed to identify foreign inclusions stemming from languages other than English as the current system is designed specifically for German and English. Given the available resources, the tool can be expanded to any language scenario.

Table 4. Sources and examples of errors

Sources of Error	**Examples**	**LANG**
Wrong POS tag	Novgorod (NN)	EN
Mixed-lingual unhyphenated compounds	**Shuttle**flug	DE
New internationalisms	Euro	EN
Abbreviations with several expansions	BIP	EN
Unreliable Google hits	Kameratelefon	EN
Inclusions from other languages	Accessoires	DE

5 Conclusions and Future Work

We have presented a tool that exploits the information listed in lexicons and published on the Web to classify English inclusions in German text on different subjects. We have shown that the size and the language distribution on the Web offer substantial resources for linguistic research as access to this corpus is publicly available.

We have demonstrated the benefit that our tool has to the work of linguists and lexicographers, as it enables them to monitor the progress of loan words and to observe the diachronic change of language over time. At the same time, we have illustrated the value of this work for research in computational linguistics, in terms of the capability of classifying foreign inclusions as well as detecting and classifying emerging lexical items by means of the abundance of documents published on the Web. Our system can be applied to new texts and domains with little computational cost and can be extended to other languages given the available lexical resources.

Future development will include extensive evaluation of our tool. Our initial aim is to use the output of our tool as a basis for annotating a gold standard which will enable us to evaluate the results in more detail. As the task of identifying and recognising foreign inclusions bears some similarities to other classification tasks such as NE recognition. We believe that a classifier based on sequence modelling may improve performance. The approaches used for these tasks, such as maximum entropy classifiers, may yield improved performance.

6 Acknowledgements

This research is supported by grants from the University of Edinburgh, Scottish Enterprise Edinburgh-Stanford Link (R36759) and the Economic and Social Research Council, UK.

References

1. Agirre, E., Martinez, D. 2000. Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of the Semantic Annotation and Intelligent Annotation workshop organised by COLING*, Saarbrücken.

2. Betz, W. 1974. Lehnwörter und Lehnprägungen im Vor- und Frühdeutschen. In Maurer, F., Rupp, H., editors, *Deutsche Wortgeschichte*, volume 1, pages 135–163. Walter de Gruyter, Berlin, New York.
3. Brants, T. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of ANLP 2000*, Seattle, WA.
4. Corr, R. 2003. *Anglicisms in German Computing Technology*. Trinity College Dublin, Computer Science Department, Dissertation.
5. Crystal, D. 2001. *Language and the Internet*. Cambridge University Press.
6. Grefenstette, G., Nioche, J. 2000. Estimation of English and non-English language use on the WWW. In *Proceedings of RIAO 2000*, pages 237–246, Paris.
7. Grefenstette, G. 1995. Comparing two language identification schemes. In *Proceedings of JADT 1995*, Rome.
8. Grefenstette, G. 1999. The WWW as a resource for example-based machine translation tasks. In *Proceedings of ASLIB 1999 Translating and the Computer*, London.
9. Grover, C., Matheson, C., Mikheev, A., Moens M. 2000. LT TTT - a flexible tokenisation tool. In *Proceedings of LREC 2000*, Athens.
10. Hohenhausen, P. 2001. Neuanglodeutsch. Zur vermeintlichen Bedrohung des Deutschen durch das Englische. In *German as a foreign language*, pages 57–87.
11. Jacquemin, C., Bush, C. 2000. Combining lexical and formatting clues for named entity acquisition from the web. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 181–189, Hong Kong.
12. Keller, F., Lapata, M. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):458–484.
13. Modjeska, N., Markert, K., Nissim, M. 2003. Using the web in machine learning for other-anaphora resolution. In *Proceedings of EMNLP 2003*, Sapporo, Japan.
14. Petrakis, H. M. 1965. *The Founders Touch. The Life of Paul Galvin of Motorola*. Chapter "The Talkies - Handie and Walkie" available online at: `http://www.batnet.com/mfwright/hthistory.html`.
15. Resnik, P. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534, College Park, Maryland.
16. Thompson, H. S., Tobin, R., McKelvie, D. 1997. *LT XML. Software API and toolkit for XML processing*. Available online at: `http://www.ltg.ed.ac.uk/software/`.
17. Viereck, W. 1986. The influence of English on German in the past and in the Federal Republic of Germany. In Viereck, W. and Bald, W.-D., editors, *English in Contact with other Languages. Studies in Honour of Broder Carstensen on the Occasion of his 60th Birthday*, pages 107–128. Budapest.
18. Volk, M. 2001. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceedings of the Corpus Linguistics Conference*, pages 601–606, Lancaster.
19. Waterman, J. T. 1991. *A History of the German Language*. Revised edition. Waveland Press, Prospects Heights, Illinois.
20. Yang, W. 1990. *Anglizismen im Deutschen: am Beispiel des Nachrichtenmagazins Der Spiegel*. Niemeyer, Tübingen.
21. Yeandle, D. 2001. Types of borrowing of Anglo-American computing terminology in German. In Davies, M. C., Flood, J. L., and Yeandle, D. N., editors, *Proper Words in Proper Places: Studies in Lexicology and Lexicography in Honour of William Jervis Jones*, pages 334–360. Stuttgarter Arbeiten zur Germanistik 400, Heinz, Stuttgart.