Automatic Detection of English Inclusions in Mixed-lingual Data with an Application to Parsing

Beatrice Alex

Doctor of Philosophy

Institute for Communicating and Collaborative Systems

School of Informatics

University of Edinburgh

2008

Abstract

The influence of English continues to grow to the extent that its expressions have begun to permeate the original forms of other languages. It has become more acceptable, and in some cases fashionable, for people to combine English phrases with their native tongue. This language mixing phenomenon typically occurs initially in conversation and subsequently in written form. In fact, there is evidence to suggest that currently at least one third of the advertising slogans used in Germany contain English words.

The expansion of the Internet, coupled with an increased availability of electronic documents in various languages, has resulted in greater attention being paid to multilingual and language independent applications. However, the automatic identification of foreign expressions, be they words or named entities, is beyond the capability of existing language identification techniques. This failure has inspired a recent growth in the development of new techniques capable of processing mixed-lingual text.

This thesis presents an annotation-free classifier designed to identify English inclusions in other languages. The classifier consists of four sequential modules being pre-processing, lexical lookup, search engine classification and post-processing. These modules collectively identify English inclusions and are robust enough to work across different languages, as is demonstrated with German and French. However, its major advantage is its annotation-free characteristics. This means that it does not need any training, a step that normally requires an annotated corpus of examples.

The English inclusion classifier presented in this thesis is the first of its type to be evaluated using real-world data. It has been shown to perform well on unseen data in both different languages and domains. Comparisons are drawn between this system and the two leading alternative classification techniques. This system compares favourably with the recently developed alternative technique of combined dictionary and n-gram based classification and is shown to have significant advantages over a trained machine learner.

This thesis demonstrates why English inclusion classification is beneficial through a series of real-world examples from different fields. It quantifies in detail the difficulty that existing parsers have in dealing with English expressions occurring in foreign language text. This is underlined by a series of experiments using both a treebank-induced and a hand-crafted grammar based German parser. It will be shown that interfacing

these parsers with the annotation-free classifier presented here results in a significant improvement in performance. It is argued that English inclusion detection is a valuable pre-processing step with many applications in a number of fields, the most significant of which are parsing, text-to-speech synthesis, machine translation and linguistics & lexicography.

Acknowledgements

There are many people who have supported me throughout the work on my thesis and to whom I would like to express my sincerest thanks. I am particularly grateful to the support and encouragement of my supervisors Claire Grover and Frank Keller who have been abundantly helpful and assisted me in all my work. I would also like to thank my examiners Ewan Klein, John Carroll and Mark Stevenson for their helpful comments. Thanks also go to Steven Clark who co-supervised me in the first year of my PhD. I am indebted to Amit Dubey and Martin Forst for assisting me in the use of their parsing tools and giving up their time to provide me with extremely helpful feedback on parsing-related work of the thesis.

I would like to acknowledge the people who have provided me with useful data, annotation or software, discussed research and gave me constructive feedback on my work. They include Mirella Lapata, Simon King, Rob Clark, Maria Milosavljevic, Malvina Nissim, Beat Pfister, Alexander Onysko, Robert Eklund, Thierry Poibeau, Thomas Segler, Ralf Steinberger and John Laffling.

I am grateful to the Economic and Social Research Council (ESRC), the Edinburgh-Stanford Link and the School of Informatics at the University of Edinburgh for awarding me with various grants which supported me during my PhD. I would also like to thank all colleagues at the School of Informatics for their kind assistance, encouragement and understanding, particularly the TXMers Barry, Mijail, Mike, Richard, Stuart and Xing.

Thanks are due also to all my friends for their support. Sasha, Ben, Anna, Gail and Fiona deserve special mention. Lastly, and most importantly, I want to thank my husband Keith and my family on whose constant encouragement and love I have relied throughout my studies.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Beatrice Alex*)

Contents

1 Introduction **1**

1.1 Related Publications and Presentations 6

2 Background and Theory **7**

2.1 Language Mixing with English 8

2.1.1 Borrowing versus Code-switching 9

2.1.2 English in German 11

2.1.3 English in French 23

2.1.4 English in Other Languages 28

2.2 Automatic Language Identification 32

2.2.1 Language Identification of Mixed-lingual Data 35

2.3 Chapter Summary 45

3 Tracking English Inclusions in German **46**

3.1 Motivation 47

3.2 Corpus Description and Preparation 48

3.2.1 Data 48

3.2.2 Annotation 48

3.2.3 Inter-annotator Agreement 51

3.2.4 Annotation Issues 53

3.3 English Inclusion Classifier: System Overview 54

3.3.1 Processing Paradigm 56

3.3.2 Pre-processing Module 56

3.3.3 Lexicon Lookup Module 57

3.3.4 Search Engine Module 60

3.3.5 Post-processing Module 63
3.3.6 Document Consistency Checking 66
3.3.7 Output 66
3.4 Evaluation and Analysis 68
3.4.1 Evaluation of the Tool Output 68
3.4.2 Evaluation of Individual System Modules 69
3.4.3 Evaluation on Unseen Data 80
3.5 Parameter Tuning Experiments 85
3.5.1 Task-based Evaluation of Different POS taggers 85
3.5.2 Task-based Evaluation of Different Search Engines 89
3.6 Machine Learning Experiments 92
3.6.1 In-domain Experiments 92
3.6.2 Cross-domain Experiments 95
3.6.3 Learning Curve 96
3.7 Chapter Summary 98

4 System Extension to a New Language **100**
4.1 Time Spent on System Extension 101
4.2 French Development and Test Data Preparation 102
4.3 System Module Conversion to French 104
4.3.1 Pre-processing Module 105
4.3.2 Lexicon Module 107
4.3.3 Search Engine Module 107
4.3.4 Post-processing Module 108
4.4 French System Evaluation 109
4.4.1 Evaluation on Test and Development Data 109
4.4.2 Evaluation of the Post-Processing Module 112
4.4.3 Consistency Checking 112
4.5 Chapter Summary 114

5 Parsing English Inclusions **116**
5.1 Related Work 117
5.2 Data Preparation 118
5.2.1 Data Sets 119

5.3 Treebank-induced Parsing Experiments 121
5.3.1 Parser 121
5.3.2 Parser Modifications 123
5.3.3 Method 125
5.3.4 Results 127
5.3.5 Error Analysis 131
5.3.6 Discussion 135
5.4 Parsing Experiments with a Hand-crafted Grammar 139
5.4.1 Parser 139
5.4.2 Parsing Modifications 141
5.4.3 Method 142
5.4.4 Results 143
5.4.5 Discussion 144
5.5 Chapter Summary 145

6 Other Potential Applications **146**
6.1 Text-to-Speech Synthesis 147
6.1.1 Pronunciation of Foreign Words 147
6.1.2 Brief Overview of a TTS System 157
6.1.3 Evaluation of TTS Synthesis 160
6.1.4 Polyglot TTS Synthesis 161
6.1.5 Strategy for Task-based Evaluation with TTS 162
6.2 Machine Translation 165
6.3 Linguistics and Lexicography 170
6.4 Chapter Summary 173

7 Conclusions and Future Work **174**
7.1 Thesis Contributions 175
7.2 Future Work 176

A Evaluation Metrics and Notation **177**
A.1 System Evaluation Metrics 177
A.2 Inter-annotator Agreement Metrics 179
A.2.1 Pairwise accuracy and F-score 179

A.2.2 Kappa Coefficient 180
A.3 Parsing Evaluation Metrics 181
A.3.1 Labelled Precision, Recall and F-score 181
A.3.2 Dependency Accuracy 181
A.3.3 Bracketing Scores 182
A.4 Statistical Tests 182
A.4.1 Chi-Square Test 183

B Guidelines for Annotation of English Inclusions **185**
B.1 Introduction 185
B.1.1 Ph.D. Project 185
B.1.2 Annotation Guidelines 185
B.1.3 Annotated Data 186
B.2 Annotation Instructions 186
B.2.1 General Instructions 186
B.2.2 Specific Instructions 187

C TIGER Tags and Labels **192**
C.1 Part of Speech Tag Set in TIGER 192
C.2 Phrase Category (Node) Labels in TIGER 194
C.3 Grammatical Function (Edge) Labels in TIGER 195

Bibliography **197**

List of Figures

2.1 Top ten internet languages 8
2.2 Orthographical variants in *Der Spiegel* between 1994 and 2000 13
2.3 Onysko's classification of the term anglicism 17
2.4 Splitting options for a mixed-lingual compound 20
2.5 Hinglish song lyrics 31
2.6 Mixed-lingual German/English analyser 38
2.7 Mixed-lingual analyser output 39
2.8 Hidden Markov Model architecture 42

3.1 System architecture of the English inclusion classifier 55
3.2 Yahoo queries with different language preferences 62
3.3 Performance increase of the corpus search module with access to increasing corpus sub-sets 76
3.4 Learning curve of a supervised machine learning classifier versus the performance of the annotation-free English inclusion classifier 97

4.1 Time spent on system extension 102
4.2 Extended system architecture 106

5.1 Example parse tree of a German TIGER sentence containing an English inclusion 118
5.2 Sentence length distribution of the inclusion set and a completely random TIGER data set 120
5.3 Parse tree with local rule probabilities 122
5.4 Tree transformation for a coordinated noun phrase rule 123
5.5 Tree transformation employed in the inclusion entity parser 126

5.6 Average relative token frequencies for sentences of equal length . . . 130
5.7 Partial parse trees of produced by the baseline parser, found in the gold standard and output by the inclusion entity model 136
5.8 Partial parse of the inclusion entity model for a false positive inclusion 137
5.9 Complete c- and f-structures for an English example sentence 140

List of Tables

2.1 German singular and plural noun inflection 21
2.2 Language origins of words in three test scripts (Marcadet *et al.*, 2005) 41
2.3 Token-based language identification error rates (in percent) for three different test scripts and methods (Marcadet *et al.*, 2005) 41

3.1 English inclusions in the German data 50
3.2 Most frequent English inclusions per domain 51
3.3 Contingency table for the English inclusion annotation 52
3.4 Difficult annotation examples 53
3.5 Most frequent homographs found in both lexicons 58
3.6 Search engine module frequencies 63
3.7 Post-processing rules 64
3.8 Performance of the English inclusion classifier on the development set 69
3.9 Evaluation of individual system modules 71
3.10 Evaluation of the corpus search module using the Wall Street Journal corpus 74
3.11 Evaluation of the corpus search module using increasing sub-sets of the Gigaword corpus 75
3.12 Evaluation of the post-processing module on the German development data 78
3.13 Evaluation of the document consistency checking step on the German development data 79
3.14 Performance of the English inclusion classifier on the German test set 81
3.15 Performance of the English inclusion classifier on the entire German data set 83

3.16 Most frequent English inclusions in the German test set of Marcadet *et al.* (2005) 84
3.17 Task-based evaluation of three POS taggers 88
3.18 Time comparison for web corpus estimation using Yahoo and Google 90
3.19 Task-based search engine evaluation 91
3.20 In-domain experiments with a trained machine learning classifier . . . 94
3.21 Cross-domain experiments with a trained machine learning classifier . 96

4.1 English inclusions in the French and German data 103
4.2 Most frequent English inclusions in the French internet data 104
4.3 Comparison of the performance of the English inclusion classifier on French and German data 110
4.4 Evaluation of the post-processing module on the French development data 113
4.5 Evaluation of the document consistency checking step on the French and German data 114

5.1 POS tags of English inclusions 125
5.2 Meaning of diacritics indicating statistical (in)significance of t-tests . 128
5.3 Baseline and perfect tagging results and results for the word-by-word and the inclusion entity models 128
5.4 Means and standard deviations of frequency profiles 131
5.5 Gold standard phrasal categories of English inclusions 131
5.6 Bracket frequency in output compared to the gold standard 133
5.7 Phrase bracketing and phrase category errors made by the baseline and inclusion entity models 134
5.8 LFG parsing results for baseline and multi-word inclusion parsing . . 143

A.1 Contingency table of gold standard and system output 178
A.2 Contingency table of two annotators 179
A.3 Agreement interpretation of κ-values 180
A.4 Contingency table of baseline and classifier 183
A.5 Critical χ^2 values and associated significance levels α 184

B.1 Amount of annotated German text per domain in tokens 186

Chapter 1

Introduction

Die Gewalt der Sprache ist nicht,
dass sie das Fremde abweist,
sondern dass sie es verschlingt.

The power of language is not
that it rejects foreign elements
but that it devours them.

JOHANN WOLFGANG VON GOETHE

Geographic language mixing is a well known phenomenon that occurs on the borders between countries, where two different languages borrow useful words, phrases or sayings from each other. However, the growth of Internet has radically changed this once localised phenomenon into virtual language mixing, as it sweeps away all physical restrictions once imposed by borders. There are now practically no limits to language mixing as an Internet page can contain text in many different languages. Given that the majority of text published on the Internet is in English, an ever increasing number of its expressions and names are appearing even in once geographically distant languages. A good example of this phenomenon is German where the number of anglicisms in the language has increased considerably in recent years (e.g. Yang, 1990; Schütte, 1996; Utzig, 2002; Androutsopoulos *et al.*, 2004; Tomaschett, 2005). There is also strong evidence for the increasing influence of English in many other languages such as French, Chinese, Japanese, Korean, Hindi and Russian. Even so, the intensity of English infiltration, as well as the impetus behind this process, can vary significantly between these languages. It has therefore become a focal point for con-

siderable research over the last few years (e.g. Kupper, 2003; Hall-Lew, 2002; Moody, 2006; Yoneoka, 2005; Kachru, 2006; Dunn, 2007). In fact, entire conferences are now dedicated to the subject area (e.g. the international conference on *Anglicisms in Europe* (2006) and the workshop on *Strategies of Integrating and Isolating Non-native Entities and Structures* to be held in February 2008).

Any form of language mixing presents a source of considerable difficulty in natural language processing (NLP) applications that automatically parse data, perform text-to-speech synthesis or translate data between languages. Current tools make the assumption that there is no language mixing, considering any input text to be monolingual, and consequently fail to process foreign language inclusions correctly. In machine translation, this failure could result in anything from direct transfer of the foreign word into processed text to a complete loss of meaning through language confusion. Linguists who study the language mixing phenomenon normally rely on painstaking manual analysis of data to draw conclusions about the occurrence of anglicisms in any given language. This is not only time-consuming and cumbersome, but also date specific. The reason for this is that languages evolve over relatively short timescales; the point at which collection of data for a corpus begins will almost certainly have different language mixing characteristics from the point at which it ends, especially if the corpus is collected over a few years. This underlines a real need for new automatic foreign language inclusion detection techniques.

Increased availability of electronic data in different languages has encouraged the NLP research community to devote greater attention to multilingual and language independent applications. However, the task of detecting foreign words and names in mixed-lingual text remains beyond the capabilities of existing automatic language identification techniques (e.g. Beesley, 1988; Dunning, 1994; Cavnar and Trenkle, 1994; Damashek, 1995; Ahmed *et al.*, 2004) which only really perform well when identifying the base language of a sentence, paragraph or document. The problem with such existing techniques is that they are based on character language models or n-gram frequencies, both of which statistically analyse typical character sequences in languages. Unfortunately, this means that they are not suited to language identification at the word level and therefore do not deal well with mixed-lingual text. Only a few research groups are active in the field of mixed-lingual language identification (Pfister and Romsdorfer, 2003; Marcadet *et al.*, 2005; Farrugia, 2005; Andersen, 2005, all

reviewed in Chapter 2). None of their algorithms have been extensively evaluated on unseen data, using instead fixed sets of data that were also used during the algorithm design phase. Some of them rely on continued human input as the language evolves either to annotate new training data or to generate new language rules. It is therefore unclear how these methods perform on unseen data from a new domain or how much effort is involved to extend them to a new language. This diminishes the benefit of these algorithms, particularly in NLP applications, given that continued human interaction is necessary to ensure accurate processing.

This thesis examines the hypothesis that it is possible to create a self-evolving system that automatically detects English inclusions in other languages with minimal linguistic expert knowledge and little ongoing maintenance. It proposes a solution combining computationally inexpensive lexicon lookup and dynamic web-search procedures that will verify and optimise its output using post-processing and consistency checking. This novel approach to English inclusion detection will then be extensively evaluated on various data sets, including unseen data in a number of domains and in two different languages. The thesis also presents extrinsic evaluation experiments to test the usefulness of English inclusion detection for parsing. It will show that by providing knowledge about automatically detected English multi-word inclusions in German to both a treebank-induced and a hand-crafted grammar parser, performance or coverage can be improved significantly. Successful demonstration of such an **English inclusion classifier** solves a significant problem faced by the NLP community in ensuring accurate and reliable output given the growing challenge of language mixing in an Internet connected world. This thesis consists of six chapters each of which examines distinct aspects of this work. They are outlined in the following paragraphs:

Chapter 2: Background and Theory presents the linguistic background and theoretical knowledge that lies behind this thesis. It first introduces the linguistic phenomenon of language mixing due to the increasing influence of English on other languages, proceeding to provide an overview of different types and frequencies of English inclusions in German, French and a few other languages. The historical background and attitudes towards the influx of anglicisms are also discussed. The chapter then reviews related work on automatic language identification and discusses four alternative approaches to mixed-lingual text analysis.

Chapter 3: Tracking English Inclusions in German describes an English inclusion classifier developed for mixed-lingual input text with German as the base language. It focuses initially on evaluation data preparation and annotation issues, subsequently providing a complete system description. The chapter also presents an evaluation of the English inclusion classifier and its components, as well as its performance on two unseen datasets. The results show that the classifier performs well on new data in different domains and compares well to another state-of-the-art mixed-lingual language identification approach. The penultimate section describes and discusses parameter tuning experiments conducted to determine the optimal settings for the classifier. Finally, the English inclusion classifier is compared to a supervised machine learner.

Chapter 4: System Extension to a New Language describes the adaptation of the classifier to process French text containing English inclusions. The aim of this chapter is to illustrate the ease with which the system can be adapted to deal with a new base language. The chapter first describes data preparation and then explains the work involved in extending various system modules. Finally, a detailed evaluation on unseen test data and a comparison of the classifier's performance across languages is presented and discussed. The results show that the English inclusion classifier not only performs well on new data in different domains but also successfully fulfils its purpose in different language scenarios.

Chapter 5: Parsing English Inclusions concentrates on applying the techniques developed in the previous two chapters to a real-world task. This chapter presents a series of experiments on English inclusions and a set of random test suites using a treebank-induced and a hand-crafted rule-based German grammar parser. The aim here is to investigate the difficulty that state-of-the-art parsers have with sentences containing foreign inclusions, thereby determining the reasons for inaccuracy by means of error analysis and identifying appropriate ways of improving parsing performance. The ultimate goal of this chapter is to highlight the oft-forgotten issue of English inclusions to researchers in the parsing community and motivate them to identify ways of dealing with inclusions by demonstrating the potential gains in parsing quality.

Chapter 6: Other Potential Applications discusses in detail a number of different fields in which automatic identification of foreign inclusions would prove beneficial, including text-to-speech synthesis, machine translation, linguistics and lexicography. The implications of applying English inclusion detection as a pre-processing step in a text-to-speech synthesiser are discussed in detail. Furthermore, a strategy is proposed for extrinsic evaluation of the inclusion classifier. This strategy is based on extensive reviews of studies dealing with production and perception of mixed-lingual speech, second language acquisition and all aspects of synthesising speech containing foreign inclusions. The chapter then rounds off with case studies on the usefulness of English inclusion detection for machine translation as well as linguistics and lexicography.

Chapter 7: Conclusions and Future Work reiterates the key lessons learnt during this program of research, summarises all the core contributions made here and examines the directions that the author believes should be taken with any subsequent body of work.

1.1 Related Publications and Presentations

The following publications are available online at: `http://www.ltg.ed.ac.uk/~balex`.

Beatrice Alex, Amit Dubey, and Frank Keller. (2007). Using Foreign Inclusion Detection to Improve Parsing Performance. In *Proceedings of EMNLP-CoNLL 2007*, Prague, Czech Republic.[1]

Beatrice Alex. (2006). English Inclusions in Mixed-lingual Data. *Draft dissertation defence presentation*, School of Informatics, University of Edinburgh, Edinburgh, UK.

Beatrice Alex. (2006). Integrating Language Knowledge Resources to Extend the English Inclusion Classifier to a New Language. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Beatrice Alex. (2005). Automatic Identification of English Inclusions in Different Languages. *Ph.D. research presentation*, Firbush, Scotland, UK.

Beatrice Alex. (2005). An Unsupervised System for Identifying English Inclusions in German Text. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005) - Student Research Workshop*, Ann Arbor, Michigan, USA.

Beatrice Alex. (2004). Classification of Foreign Inclusions in Mixed-lingual Data. *Research Proposal*, School of Informatics, University of Edinburgh, Edinburgh, UK.

Beatrice Alex. (2004). Identification of Foreign Inclusions in Mixed-lingual Data. In *IGK Summer School*, Firbush, Scotland, UK.

Beatrice Alex, and Claire Grover. (2004). An XML-based Tool for Tracking English Inclusions in German Text. In *PAPILLON 2004: Workshop on Multilingual Lexical Databases*, Grenoble, France.[2]

Beatrice Alex. (2004). Mixed-lingual Entity Recognition. *Informatics Jamboree*, School of Informatics, University of Edinburgh, Edinburgh, UK.

---

[1] Authors' contributions: BA prepared the data, carried out all the experiments and analysed the results, AD modified the parser used in the experiments, and FK provided feedback. All authors contributed in the writing of the paper.

[2] Authors' contributions: BA prepared the data, developed the classifier, and wrote the paper. CG gave supervisory feedback on the paper.

Chapter 2

Background and Theory

This chapter presents the background and theory behind the work presented in this thesis: (1) the phenomenon of anglicisms occurring in other languages, and (2) previous work aiming to identify language portions in mixed-lingual text. The chapter first describes the phenomenon of **language mixing** which is the ultimate cause for the research conducted as part of this thesis project. A specific type of language mixing is discussed, namely that caused by the growing influence of English on other languages, resulting in an increasing use of anglicisms. Particularly, the phenomena of **borrowing** and **code-switching**, related manifestations of such language contact, are discussed in detail. Subsequently, this chapter examines the types of anglicisms occurring in German, French and other languages. The historical background of this language mixing phenomenon, an underlying linguistic theory of anglicisms, the types and frequency of English forms found in those languages and attitudes towards their usage are presented. The chapter continues with a review of previous and related work on automatic **language identification** (LID) with particular focus on **mixed-lingual LID**. Note that all evaluation metrics and notations used by reviewed work and applied in the experiments which are presented in this thesis are outlined in Appendix A.

2.1 Language Mixing with English

Languages do not exist in isolation and have some degree of influence on each other when they come into contact, for example, at the border of two countries where different languages are spoken, due to migration of people to other countries or in multi-ethnic societies. Such strong physical contact between two speech communities is not the only cause for language mixing. In the past decade, there has been an increased focus on studying the linguistic effects of language contact due to rapid globalisation and the increasing popularity of the internet. Internet usage has greatly influenced the course of different languages. Computer-related activities and communication technologies lead to new models of communication (Tosi, 2001). English plays a significant role as the prevailing language of internet communication. This is illustrated in Figure 2.1 which shows the estimated number of individual language speakers using the Web at the beginning of 2007.[1]

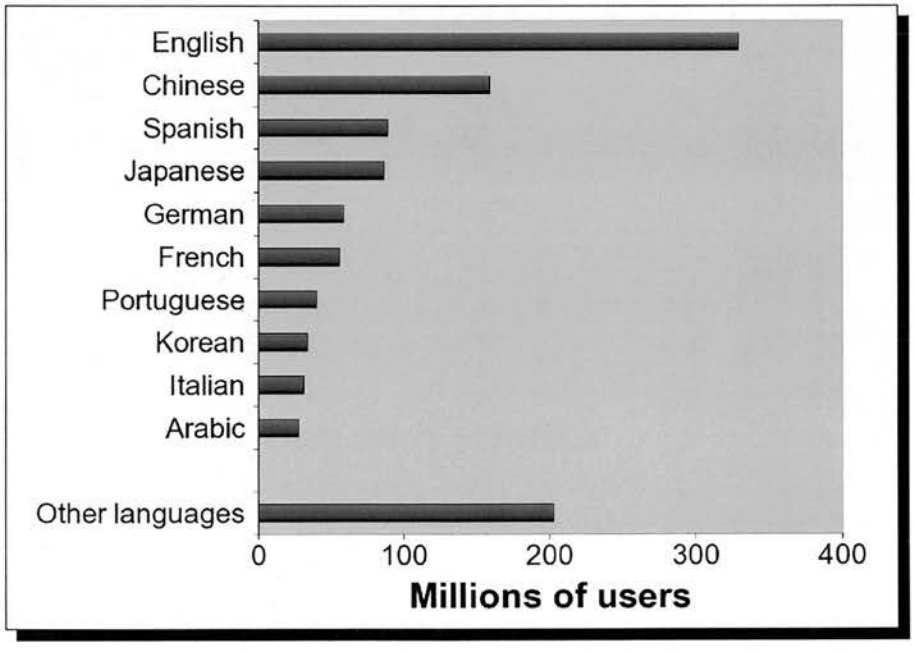


Figure 2.1: Top ten internet languages: estimated number of users per language (Internet World Stats, 2007).

[1] Statistics taken from *Internet World Stats* on June 13, 2007: `http://www.internetworldstats.com/stats7.htm`

The number of internet users has increased nearly 71-fold from an estimated 16 million in December 1995 to 1,133 million in June 2007.[2] While the overall presence of English on the Web has declined during that time period, Figure 2.1 clearly shows that the internet is still English-dominated. The preponderance of English on the Web results in virtual language contact and concomitantly in language mixing. Two extensively studied phenomena that are the results of this type of language mixing are language borrowing and code-switching which are discussed next.

2.1.1 Borrowing versus Code-switching

Both borrowing and code-switching are language contact phenomena that are defined and delineated by many linguists (e.g. Sankoff and Poplack, 1984; McClure and McClure, 1988; Myers-Scotton, 1993; Poplack, 1993; Gardner-Chloros, 1995; King, 2000; Muysken, 2000; Callahan, 2004; Onysko, 2007, to name but a few). While some of them make a clear distinction between both phenomena based on certain criteria, others use the term **code-mixing** to include any type of language mixing as it is not always clear where borrowing stops and code-switching begins. The criteria that are used to classify foreign inclusions as either borrowings or code-switches are: the number of lexical units, morpho-syntactic integration into the receiver language, and usage frequency. According to these three criteria, the majority of linguists define borrowings as single lexical units from a source language that can be structurally and sometimes phonologically modified when embedded into the receiver language. There is also a tendency to regard foreign expressions as borrowings if they are frequently used (e.g. Sankoff and Poplack, 1984; Myers-Scotton, 1993; Muysken, 2000). Proper nouns, including names of organisations, companies and brands can generally be classified as borrowings.

Conversely, code-switches are defined as multi-word units from a source language that retain their structure and pronunciation when embedded in the receiver language (e.g. Onysko, 2006). Regarded as distinct from the surrounding text, Myers-Scotton (1993) appropriately calls them "embedded language islands". In terms of their usage frequency, code-switches are often thought of as single occurrences of foreign elements. Typically, code-switching is regarded as a symptom of bi- and multilingual

[2]Statistics taken from *All About Market Research* on June 13, 2007: `http://www.allaboutmarketresearch.com/internet.htm`

speakers in diverse discourse situations. However, Onysko (2006) argues that code-switching in written language, yet less extensively investigated, is a phenomenon that does exist. Code-switching includes English phrasal and clausal segments occurring, for example, in the German text as a result of English developing into a global means of communication and its increasing impact on German. One justification for this theory is the high degree of latent bilingualism among German native speakers. The author of a written piece of text primarily acts as a mediator of the code-switch with the aim to induce a receptive code-switch in the reader.

While code-switching involves the insertion of foreign elements into the receiver language, borrowing leads to the foreign element being entered into the receiver language speaker's lexicon (Muysken, 2000). However, this well-defined differentiation can be problematic as certain foreign inclusions are difficult to classify according to these criteria. Onysko (2006) points out that the definitions do not account for single-word code-switches and multi-word borrowings both of which occur. He lists examples of English inclusions found in German newspaper text such as *trial and error* or *Books on Demand* which, although multi-word English inclusions, appear as lexical elements in the text and therefore follow the notion of borrowings. Moreover, Callahan (2004) suggests that foreign, single-word company and brand names could be regarded as code-switches as they are not subject to structural or phonological adaptation. Differentiating between borrowing and code-switching in spoken language can also be ambiguous if the pronunciation of the foreign inclusion is imperfect (Poplack, 1988).

It can therefore be concluded that there is no clear-cut distinction between language borrowing and code-switching. In fact, some linguists prefer to define both related manifestations of language contact as a continuum ranging from borrowing to code-switching with non-canonical cases in between (e.g. Boyd, 1993; Clyne, 1993). A differentiation is often dependent on the given situation of language contact. Onysko (2006), for example, carried out a corpus analysis investigating the occurrence of English inclusions in German newspaper text. He found that, with some exceptions, the majority of inclusions account for either borrowings, single-word inclusions that largely follow the morpho-syntactic conventions of German, or code-switching, multi-word inclusions governed by English syntactic rules. The impact of English on German, French and other languages, particularly in written language, is examined in more detail in the following sections.

2.1.2 English in German

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 in German. Androutsopoulos *et al.* (2004) show that, after 2000, the number of English slogans in German advertising amounted to 30%, compared to the 1980s when only 3% were English, a 10-fold increase in only 20 years. In some domains such as IT or clothing, the percentage of English slogans reaches of 50%. However, borrowing from English is not a new development.

2.1.2.1 Historical Background

Language contact with English dates as far back as the 8th century but was, at that time, limited to the Northern regions of today's Germany and mainly occurred in the domains of religion and trade (Viereck, 1984). First anglicisms like the word *Boot* (boat) appeared in German during the Middle Ages as a result of emerging trade with English merchants in the Rhineland as well as trade relations between Britain and the Hanseatic League (Huffman, 1998; Viereck, 1984). However, the number of anglicisms found in German at that time is relatively small, amounting to 31 at most as suggested by some studies (cf. Hilgendorf, 2007). With the Industrial Revolution in the 18th century, English became more and more popular in German-speaking territories and its growing influence eventually presented a challenge to the well-established status of French which used to signal social prestige (Gentsch, 1994). However, English had an even stronger influence on German during the 19th and 20th centuries. Hermann Dunger emerged as one of the first critics of this trend (Hilgendorf, 2007). At the end of the 19th century, he published a dictionary of superfluous germanised foreign words and protested severely against the influx of English into the German language (Dunger, 1882, 1909, both reprinted in 1989).

While British English was the main source of borrowing before World War II, the establishment of the USA as a global power resulted in a concomitant influx of American English expressions into German (Hilgendorf, 1996). This development was further amplified by technological advances such as the invention of the internet as well as increasing globalisation. Towards the end of the 20th century and at the beginning of the 21 century, linguists recorded an enormous increase in the number of anglicisms

entering German in many domains of life. In East Germany, this development gained momentum only after the reunification in 1990. During the preceding 40 years of socialism, anglicisms were disapproved for epitomising the manipulating influence of the enemy of the people. According to Fink *et al.* (1997), the difference in the frequency usage of anglicisms between West and East Germans almost completely disappeared towards the end of the 1990s.

2.1.2.2 Frequency of English Inclusions

Concerning the frequency of anglicisms occurring in German, in the last 50 years, a whole range of corpus studies have been conducted by numerous linguists on different types of corpora and domains. Carstensen *et al.* (1972), for example, found an average of 10 anglicisms per page when examining a German newspaper published in 1971. In earlier studies carried out in 1963 and 1965, he had estimated only one or two anglicisms per newspaper page (Carstensen, 1963, 1965). However, Busse (1993) shows that even in the 1980s only relatively few anglicisms were actually listed in West and East German versions of the German dictionary *Duden*.

A large number of diachronic studies were conducted to examine the influence of English on German over time. Yang (1990) reports an increase from 2.93 to 4.39 anglicisms per page of the German magazine *Der Spiegel* between 1950 and 1980. Schütte (1996) shows that the percentage of anglicisms present in advertisement sections of three German magazines increased from 1.1% to 6.7% between 1951 and 1991. Utzig (2002) conducted a similar analysis but specifically focusing on determining the proportion of English job titles and descriptions in newspaper job adverts. He finds that the proportion of anglicisms increases from 1.6% to 35.6% between the years 1960 and 1999. Therefore, at the turn of the last century, more than one third of job titles and descriptions advertised in German newspapers are in English. This percentage is likely to have increased even further in recent years as English expressions are becoming more and more popular in German. Androutsopoulos *et al.* (2004) detects a 10-fold increase in the number of English slogans in German advertisements from the 1980s to after 2000. Tomaschett (2005) carefully examined advertisement sections of three Swiss newspapers and found that between 1989 and 2005 their proportion of anglicisms has risen from 2.9% to 8.9%, a 3-fold increase. One fifth of anglicisms refer to internet-related vocabulary signalling that the rise of the internet is a crucial

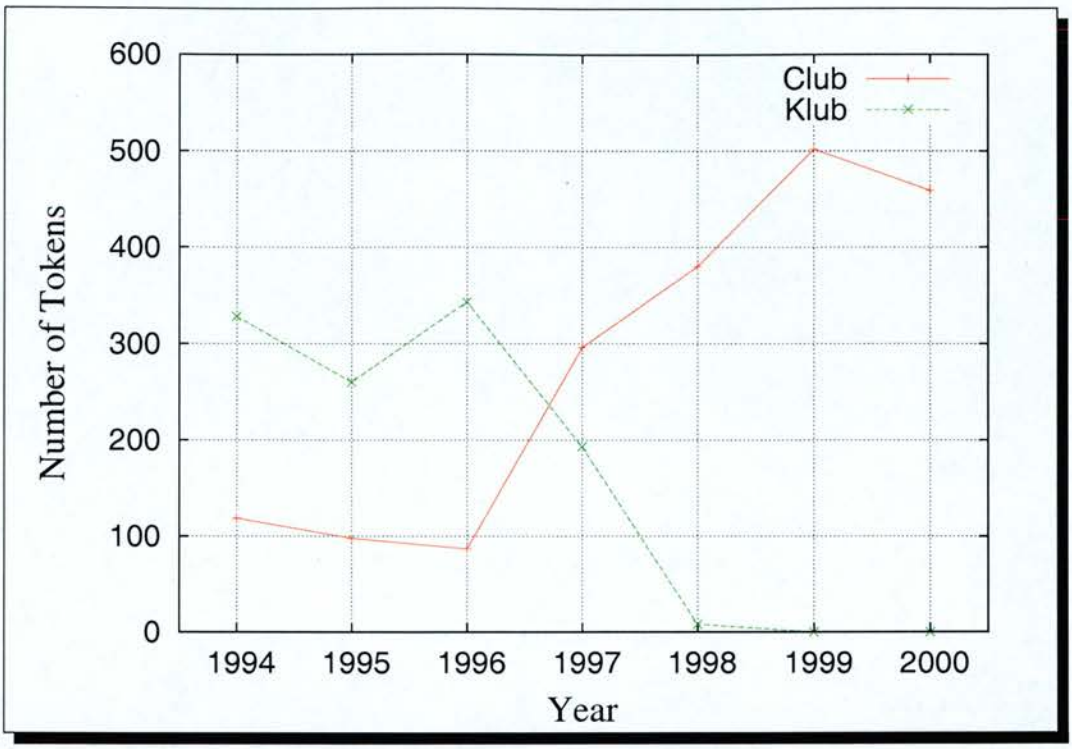


Figure 2.2: Frequency of orthographical variants *Club* and *Klub* in *Der Spiegel* between 1994 and 2000 (Onysko, 2007).

factor in the increasing influence of English on German. Tomaschett found that there is no substantial difference in the frequency of anglicisms across different newspapers but that figures vary considerably across domains. He observes that non-commercial domains like obituaries and official announcements contain far fewer anglicisms than commercial adds. However, even their initially small proportion of anglicisms in 1989 has increased considerably over the time period that was investigated. A most recently published diachronic corpus study on anglicisms in written German is that of Onysko (2007). His analysis shows that, apart from the capitalisation of nominal borrowings and variations in the concatenation of compounds, many anglicisms used in *Der Spiegel* follow the principle of zero substitution on the graphemic level as illustrated by one example in Figure 2.2. The frequency of the anglicism *Club* increased dramatically in the *Der Spiegel* issues published between 1994 and 2000. Conversely, the assimilated variant *Klub*, which was still popular up until 1996, was no longer used in 1999 and 2000.

Onysko analysed the articles of *Der Spiegel* published in 2000 in greater detail. He reports that 1.11% of all tokens and 5.80% of all types are made up of anglicisms with a corresponding type-token-ratio of 0.29. The basis for this corpus study is a formal definition of the term anglicism which is explained further in Section 2.1.2.4. This largely overlaps with the types of English inclusions recognised by the English inclusion classifier described and evaluated in Chapters 3 and 4. In German newspaper text published in the *Frankfurter Allgemeine Zeitung*, the evaluation data used to assess the classifier, the percentage of anglicisms varies depending on the domain ranging from 6.4% of tokens for articles in the domain of IT to 0.3% of tokens for articles on topics related to the European Union (see Table 3.1 in Chapter 3).

Regarding the parts-of-speech of anglicisms found in German, several studies that analysed this aspect using corpora in various domains established that of all anglicisms found, nouns are the most frequent ones, accounting for more than 90% of tokens (Yang, 1990; Yeandle, 2001; Corr, 2003; Chang, 2005). Anglicisms representing other parts of speech are relatively infrequently used. For example, in Chang's analysis of a computer magazine, 3% of all anglicisms were adjectives and 0.5% verbs.

The main focus of this thesis is automatic English inclusion detection in written text. However, anglicisms often enter a language in spoken forms first, before they appear in written language. For example, Viereck (1980) found in a comprehension study that often subjects only understood certain anglicisms that were presented to them once they were pronounced to them by the experimenter. In general, studies on the frequency of anglicisms in spoken German as well as their pronunciation and perception by German speakers are relatively limited compared to the plethora of anglicism research based on written German text. This is mainly related to the availability of appropriate speech corpora but also the fact that any production and perception studies are influenced by a series of factors that are difficult to control for, including the age and the language skills of a speaker or listener (see Section 6.1.1.1). Several production experiments established that the integration of anglicisms and English sounds into German pronunciation patterns is highly correlated with a subject's age (e.g. Greisbach, 2003) and their ability to speak English (e.g. Fink, 1980). Further influential factors are the origin of the anglicism (i.e. British or American English), its orthographic integration into the receiver language and the popularity of the expression (Busse, 1994). Fink (1980) also found that pronunciation patterns of anglicisms often do not corre-

spond to their official transcriptions listed in German dictionaries. A detailed study by Glahn (2000) examining the use of anglicisms in broadcasts of two German public TV channels shows that the frequency of anglicisms varies across programmes and is highest in adverts, with an anglicism occurring on average every 23 seconds. Whereas previous research attempted to classify the pronunciation of full-word anglicisms as a whole rather than examining their individual phones, Glahn's study also provides a list of English sounds contained in anglicisms along with their pronunciations in the corpus. A recent comprehensive production and perception study of English sounds in German is that of Abresch (2007). Besides aforementioned conclusions related to the age and English language skills of subjects, she also found that English xenophones, i.e. phones that do not exist in German (see also Section 6.1.1.1), are pronounced more authentically when occurring in proper names than in other types of anglicisms. Moreover, Abresch found that subjects with a good knowledge of English nativise certain English xenophones in German sentences, even though they are able to pronounce them authentically in English contexts.

Studies on the influx of anglicisms in written and spoken German, although they examine this type of language mixing from different angles and with various definitions of anglicisms, all point to the fact that the frequency of English words and expressions in German is increasing. As a result, there is frequent exposure to German documents containing English names and expressions. This growing 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 has even appeared on Germany's political agenda (Hohenhausen, 2001).

2.1.2.3 Attitudes towards English Inclusions

Attitudes towards the influx of English expressions into written and spoken German are relatively complex and often contradictory. While some linguists view this language mixing phenomenon positively in terms of linguistic creativity (e.g. Carstensen, 1986; Görlach, 1994), language purists are warning of the decay and death of German (e.g. Weinrich, 1984; Moser, 1985). The Association of the German Language (Verein Deutscher Sprache e.V.)[3] is also extremely critical of this linguistic development. It

[3]http://www.vds-ev.de

advocates resisting the use of superfluous anglicisms in German and provides a list of alternative German expressions to use instead (Junker, 2006). The association argues that the anglicisation of German results in the exclusion of people without the necessary proficiency of English and calls the use of anglicisms "show-off" behaviour.

However, such critics are generally in the minority and positive attitudes to the occurrence of anglicisms in German clearly prevail. For example, Prof. Dr. Rudolf Hohberg, Director of the Society for the German language (Gesellschaft für Deutsche Sprache), does not believe anglicisms to be a threat to the German language (quoted from an interview by Link, 2004). He views language as an economic system where superfluous words do not enter a lexicon by default (Hoberg, 2000). Hoberg also quotes the results of a survey carried out by the Institute for German Language (Institut für Deutsche Sprache) in 1997 which found that only a quarter of Germans believe that this language mixing phenomenon is a cause for concern (Stickel, 1999). In a similar fashion, other linguists perceive this linguistic development as an indication of the high prestige English has for Germans and the fact that English, being the first foreign language taught at school, is spoken by the majority of the population (Hedderich, 2003; Harris, 2003). Berns (1992) explains the motivation behind using anglicisms in the legal domain, for example, as an attempt at absolute precision in naming concepts, a typical characteristic of legal writing. Hilgendorf (2007) argues that German native speakers switch to English as the default language when interacting with English native interlocutors, even if the German language skills of the latter are superior. As commented on by Harris (2003), some linguists go as far as interpreting the eagerness of Germans to speak English as a reaction to the xenophobia of the fascist era (Clyne, 1997; Zimmer, 1997; Busse and Görlach, 2002) or as an expression of their favourable attitude towards European unity (Graddol, 1997). Eckert *et al.* (2004) suggest that the influence of English on German, a by-product of globalisation, does not result in language death.

Whether accepted by linguists and native speakers or not, it is evident that the influx of English expressions into German has stimulated the interests of linguists and lexicographers. For them, an automatic classifier of foreign inclusions would prove a valuable tool as lexical resources need to be updated to reflect this trend. This language mixing phenomenon must also be dealt with by developers of NLP applications.

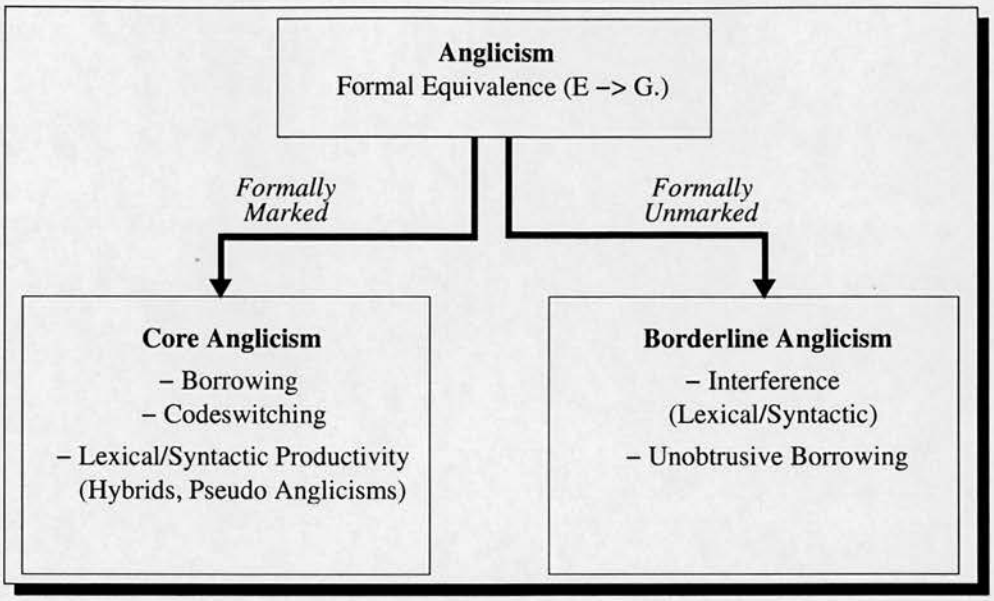


Figure 2.3: Classification of the term anglicism by Onysko (2007).

2.1.2.4 Typology of English Inclusions

There has been a plethora of research on the occurrence of anglicisms in German (and other languages) often with varying definitions of the term anglicism. The aim of this thesis is to develop an automatic English inclusion classifier that is able to identify English forms in other languages. Onysko's (2007) formal definition of anglicism visualised in Figure 2.3 serves as a suitable framework for this work.

Onysko treats the concept anglicism as a hypernym of all English forms occurring in German: borrowing, code-switching, hybrid forms, pseudo-anglicisms as well as interference and unobtrusive borrowing. Essentially, core anglicisms are, with some exceptions that are explained later, the forms that the English inclusion classifier is able to recognise. Interference, i.e. semantic and functional transfer on lexical, semantic, and pragmatic levels as a result of formal similarity of source and receiver language units like *realisieren* (to become aware of)[4], and unobtrusive borrowings like *Keks* (biscuit, from cakes) are not recognised by the classifier as they are formally unmarked in German. Onysko's definition of anglicisms is also a preferable theoretic

[4]In German, the verb *realisieren* used only to be used im the sense of *to carry out*, or *to put into practice*. Because of its similarity to the English verb *realise*, it has adopted a new sense, as in *to become aware of sth.*

framework for the work presented in this thesis as it excludes all semantic borrowing, i.e. loan substitutions (or loan coinage) which are an integral part of other definitions (e.g. Betz, 1936; Haugen, 1950; Duckworth, 1977; Carstensen, 1979; Kirkness, 1984; Yang, 1990).[5] One type of loan substitutions are loan translations, like the German word *Familienplanung* which means the same as the English expression *family planning*. Onysko (2007) views loan substitutions as conceptual transmissions without source language form and does not regard them as anglicisms but as language inherent creations. The reason is that the actual proof of possible conceptual influence is doubtful given that it is dependent on exact etymological evidence.

Currently, the English inclusion classifier is designed to recognise but not distinguish between the following types of anglicisms:

- Borrowings: *Business*, *Event*, *Software*
- Code-switching: *real big french guy*, *Gentlemen's Agreement*, *nothing at all*
- English morphemes in hyphenated hybrid forms: *Airline*-Aktien (airline share), *Computer*-Maus (computer mouse), *Online*-Dienst (online service)
- Pseudo-anglicisms: *Beamer* (video projector), *Handy* (mobile phone), *Oldtimer* (vintage car)

Before each type is discussed more extensively, it should be clarified that this thesis refers to the term anglicism without differentiating between British or American English. Borrowing and code-switching were explained in detail in Section 2.1.1 and are relatively straightforward to understand given that they involve clear language changes between words in the text. **Interlingual homographs** are a particular type of phenomenon that requires further explanation. They are forms that exist in different languages but not necessarily with the same pronunciation or semantics. For example, the German noun *Station* (hospital ward) and the English noun *Station* (as in *Space Station Crew*) belong to this category. While the large majority of interlingual homographs in German text refer to the German variant, there are exceptions. Software that disambiguates the language origin of interlingual homographs would require a

[5]The sub-classification and naming convention for these definitions vary. For a summary of these and other definitions see Corr (2003).

complex understanding of the semantics of the sentence itself. This necessitates some level of deeper semantic processing to be implemented behind the software such that it understands the overall meaning of a sentence. Developing such software is beyond the scope of this thesis, however, often interlingual homographs can be disambiguated based on their surrounding context. As the main language identification components of the English inclusion classifier described in detail in Chapter 3 are token-based, a final post-processing module was implemented to resolve ambiguities from the context.

Two important linguistic processes in German are **compounding** and **inflection** which need to be considered as English inclusions are also affected by them. Both phenomena result in the formation of hybrid, or mixed-lingual forms, in this case specifically tokens made up of English and German morphemes. Compounding is a very productive process in German. It involves concatenating two or more words to form one single orthographic new word. Combining a German word with other words in a compound can result in a virtually unlimited amount of different word forms. For example, the longest compound found in the German corpus used for the experiments in Chapter 3 is *Gruppenfreistellungsverordnungen* (group exemption regulations).

In German text, English noun compounds conform to the German compounding process and are generally concatenated with or without a hyphen. For example, *cash flow* becomes either *Cashflow* or *Cash-Flow*. Generally, the unhyphenated version is preferred unless the compound becomes too complex to read. English adjective plus noun inclusions such as *high school* are either simply capitalised (*High School*) or concatenated into one token (*Highschool*). With the increasing influence English is having on German, numerous mixed-lingual compounds have entered German, e.g. *Langzeitcrew* (long-term crew), *Keyboardtasche* (keyboard bag), *Backupdatei* (backup file). There is a tendency to hyphenate such mixed-lingual expressions in order to facilitate differentiation between individual word elements. This often results in several ways of spelling a mixed-lingual compound, e.g. *Backup-Datei*, *Back-up-Datei* or *Backupdatei*. Currently, the English inclusion classifier is designed to recognise English forms within hyphenated or non-concatenated mixed-lingual compounds.

The identification of English inclusions within unhyphenated mixed-lingual compounds requires deeper morphological analysis. A natural approach to dealing with such instances is to treat them as a concatenation of their components. In future work, the aim is to apply compound splitting techniques tested on German compounds and

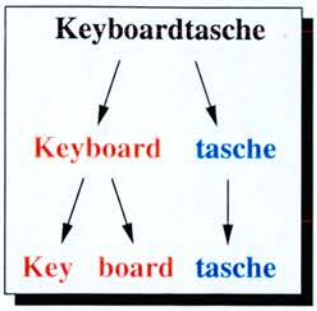


Figure 2.4: Splitting options for the mixed-lingual compound *Keyboardtasche*.

used for MT (Koehn and Knight, 2003), information retrieval (Monz and de Rijke, 2001) and speech recognition (Larson *et al.*, 2000). This work will investigate whether the algorithm that breaks up compounds into known words and fillers between words can be extended to determine English inclusions at the same time. This will be based on exhaustive recursive search of sub-strings in an English and a German lexicon in order to determine possible splitting options, as illustrated in Figure 2.4. Additional information such as frequency and POS tags of possible components can then be used to determine the preferred splitting option.

Inflection, the second linguistic process resulting in hybrid forms, follows very standardised rules in German which differ depending on the gender, the ending of a given noun and case marking. Unlike in German, English nouns do not have different grammatical gender. When used in German, they are assigned either masculine, feminine or neuter gender. This gender assignment is based on several conditions including morphological analogy, lexical similarity, latent or hidden semantic analogy or group analogy to the German equivalent. The natural gender or the number of syllables can also play a role, and anglicisms made up of verb and particle like *Make-up* always take either masculine or neuter gender (Yang, 1990).

English nouns used in German are declined according to different inflection classes which are defined based on the endings of the genitive singular and nominative plural analogous to the declension of German nouns. For example, a word belongs to the class `s/s` if both its genitive singular and nominative plural endings are *-s* (see Table 2.1). The majority of English nouns follow the declensions highlighted in light grey. While most feminine English nouns are declined according to the inflection class `ø/s`, the

Case	Singular			Plural				
Nominative	ø	ø	ø	s	ø	e	en	er
Genitive	(e)s	ø	en	s	ø	e	en	er
Dative	ø	ø	en	s	(n)	en	en	ern
Accusative	ø	ø	en	s	ø	e	en	er
Class	(e)s	ø	en	s	ø	e	en	er

Table 2.1: Inflection classes for singular and plural nouns in German.

majority of masculine and neuter English nouns fall into the class `s/s`. English nouns ending in *-er* like *Cluster* follow the null plural declension ø and receive the ending *-n* in the dative case. The plural ending of any noun stem ending in *-y* is *-s* (*Babys* meaning babies). A small number of English nouns have two plural declensions such as *Byte* (*-s*, *-ø*). There are also some nouns that do not have a plural (*Fairness*, *Publicity*).

This analysis shows that the number of different types of inflection for English inclusions nouns is relatively limited and largely predictable. For that reason, the English inclusion classifier is currently not set up to split such tokens into the English morpheme and the German ending. However, in future, it is feasible to extend the classifier to consider word stems as well. For example, the noun *Clustern* is unlikely to be listed in a German or English lexicon. However, a lookup of the string with all possible inflection endings removed will return an entry for *Cluster* from the English lexicon. If more than one ending can be removed and more than one string is found in the lexicon, the longest match will be considered. Although this method alone will not facilitate case differentiation, it is expected to produce good results in determining whether a stem is an English inclusion or not.

Another language contact phenomenon to be considered when analysing mixed-lingual data is the occurrence of **pseudo-anglicisms**. They are neologisms occurring in the receiver language which, although made up of one or more English lexical items, are sometimes unknown to or unrecognisable by English native speakers (Onysko, 2007). Many pseudo-anglicisms are of genuine English words which refer to something completely different when embedded in the receiver language, e.g. *Body Bag* (backpack), *Evergreen* (golden oldie), *Handy* (mobile phone), *Peeling* (body scrub) or

Wellness (spa). Native German speakers are not always aware of this and may believe that they are genuine anglicisms that can be used with the same meaning in English. Other pseudo-anglicisms, which are made up of one or several English morphemes, do not actually exist in English, e.g. *Beamer* (video projector), *DJane* (female DJ), *Dogwalk* (catwalk at a dog show) *Fitnessstudio* (gym) or *Pullunder* (sleeveless pullover). The examples show that pseudo-anglicisms represent the productive use of English by German native speakers. In many cases, the connection between the pseudo-anglicism and its source language item(s) is obscure. Although linguists disagree on whether pseudo-anglicisms can be classed as borrowings, it is clear that such instances would not exist in the receiver language if they had not been derived from lexical items in the source language. For example, Onysko (2007) considers them as anglicisms, but not borrowings, as they are made up of English forms but are equally the result of receiver language inherent creation. Their frequent occurrence, however, indicates the widespread influence of English in German-speaking territories. With respect to automatic language identification, pseudo-anglicisms are treated as English inclusions.

The final aspect of language mixing taken into account is the rising number of English **proper names** in German caused by the increasingly international nature of the working environment. English names of companies like *Google* or *Germanwings*, organisations like *Greenpeace* or *Fertility Center Hamburg*, events like *Fullmoon Festival* or *Love Parade* and band names like *Fury in the Slaughterhouse* or *Absolute Beginner* appear frequently in German. Such English proper names are manifestations of language contact whereby a specific concept is either transferred with its name into the receiver language or inherently created in the receiver language using English forms (Onysko, 2007). In terms of the classification of contact types, proper names appear as borrowings even though they are not consistently recognised as anglicisms by linguists. For example, Yang (1990) regards proper names, and citations related to English speaking countries as a sub-class of anglicisms. Busse (1993) also considers English proper names as anglicisms. In corpus studies, they are sometimes excluded (e.g. Onysko, 2007; Tautenhahn, 1998; Gentsch, 1994; Yang, 1990) but often included (e.g. Abresch, 2007; Hilgendorf, 2007; Corr, 2003; Hedderich, 2003; Berns, 1992; Galinsky, 1980; Koekkoek, 1958) in the analysis. Such studies tend to be limited to organisation, event, brand and product names and such like, but exclude person and location names. In this thesis, the same distinction is made and only English names of

organisations, events, brands and similar are treated as English inclusions. Identifying the language origin of person and location names is considered a task that is beyond the scope of this thesis. Moreover, person names can be used across many languages, though their pronunciation can differ. While this issue is relevant to text-to-speech (TTS) processing where it is necessary to differentiate between different pronunciation variants, language identification of person and location names is not vital to other NLP processing applications such as machine translation and corpus analysis tools for lexicographers.

2.1.3 English in French

Language mixing as a result of increasing numbers of English words and expressions appearing in another language is not limited to German. The occurrence of anglicisms and pseudo-anglicisms in French is not a new phenomenon either. One well known anglicism in French is the word *weekend* which was borrowed from English at the beginning of the 20^{th} century. However, with growing internationalisation, the influx of English expressions into the French language has taken on a different dimension in recent years. Despite serious efforts from the French government in the 1990s, which tried to restrict this trend by introducing new French words to replace already prevalent anglicisms, the French media often do not object to the use of anglicisms. This is particularly the case when a French term has not yet been invented or when a specific English term is more modern and therefore more popular than its French equivalent (Rollason, 2005; Nicholls, 2003). Moreover, the use of anglicisms in advertising, television and radio broadcasts is targeted specifically at a young audience which is heavily influenced by Anglo-American culture. Furthermore, the prominence of English on the internet has a large impact on the French language. The following sentence, taken from an online article published by ZDNet France (Dumont, 2005), contains some examples of English inclusions in French:

(1) Tous les **e-mails** entrants, qui ne seront pas dûment authentifiés par **Sender ID**, seront considérés automatiquement comme du **spam**.
Translation: *All incoming emails which will not be duly authenticated by Sender ID, will be automatically considered as spam.*

Despite extensive language planning and protection advocated by the French government, English inclusions appear regularly in various domains of French society, including the media, business and youth culture. Many linguists attribute the attractiveness of anglicisms to their concision, the law of least effort, the lack of similar terms in the receiver language, and psycho-social factors (e.g. Sokol, 2000; Tattersall, 2003). There is no doubt that the internet plays an important role in the increasing influx of anglicisms in French and other languages as they come into virtual contact with English, the prevailing language of information published in web pages and online documents.

2.1.3.1 Historical Background

Historically, language mixing between French and English has been a reciprocal process with French enriching English and vice versa. For example, after the Normans conquered England in 1066, a large number of French words entered the English language. This process was reversed in the middle of the 18th century (Tattersall, 2003). Rollason (2005) argues that some degree of such two-way, cross-lingual contamination is inevitable between neighbours as close as Britain and France. Interesting manifestations of this cross-lingual borrowing are evident in the existence of anglicisms like *flirter* and *gadget* in French from the English *to flirt* and *gadget* which linguists in turn suspect to be borrowings from the French *fleureter* (to talk sweet nonsense) and *gâchette* (catchpiece of a mechanism).[6]

Similar to German, the oldest anglicisms in French refer to trade and seafaring expressions. Before the 17th century, the occurrence of English terms in French was relatively limited (Guiraud, 1965). From that point onwards, France borrowed increasing numbers of English expressions as a result of England's growing political and economic status. This phenomenon intensified in the 19th century with the influx of English terms related to science and technology (Sokol, 2000). The occurrence of anglicisms in French accelerated even more in the 20th century resulting in the coinage of the term **Franglais** (Etiemble, 1964) and extensive criticism by language purists, including the Académie Française, an elite French institution.

This language development also resulted in the passage of two laws: the Bas-Lauriol law in 1975, imposing compulsory but non-exclusive usage of French in partic-

[6] `http://www.etymonline.com`

ular areas of French society, and the Toubon law in 1994, an extension of the preceding law, aimed at limiting the influx of English throughout society and introducing French equivalents to replace new and existing anglicisms (Tattersall, 2003; Nicholls, 2003). This legislation led to a series of lawsuits against companies such as Quick in 1984 and Body Shop in 1995 for the usage of English menu items, brands and product names (Sokol, 2000; Tattersall, 2003). This resulted in hefty fines for the culprits and an order to translate their English menu items and product names into French. The extent to which these rigorous measures have limited borrowing from English is debatable.

Tattersall (2003) suggests that the lack of available French equivalents to fast emerging anglicisms or the abundance of proposed French equivalents for the same English concept (e.g. *navigateur*, *explorateur*, *fureteur*, *lectoir*, *feuilleteur* and *broutage* all referring to the English word *browser*) can contribute to the establishment of certain anglicisms in French. Tattersall also regards French as more rigid than English, and believes the concision and flexibility of English to be key factors in the attractiveness of anglicisms. Currently, the primary source of anglicisms in French is American English. This is due to the super-power status of the USA in the world and is fuelled further by the omnipresence of the internet, a largely English language medium.

2.1.3.2 Frequency of English Inclusions

Contemporary French writing and journalism are permeated with words and phrases derived from English. For example, Rollason (2005) cites a corpus analysis quoted in Laroch-Claire (2004) which found approximately one new anglicism in every three pages of *Non merci, Oncle Sam!* (Mamère and Warin, 1999), a book which is dominated by anti-American polemics of its two authors, Noël Mamère, a politician of the Green party, and Olivier Warin, a TV journalist. Several linguists have carried out comparative studies between the number of anglicisms occurring in French and German. Zimmer (1997) found that, for a set of 100 computing terms, the native terminology of German only amounted to 57%, almost 30% lower than that of French (86%). Consequently, appropriate French translations for English terms were preferred. Plümer (2000) shows that there are 9% more anglicisms in German written and spoken media language than in French. Furthermore, Kupper (2003) determined that the proportion of anglicisms appearing in French newspaper advertisements tripled from 7% to 21% between 1976 and 2001. She found the same three-fold increase in the proportion of

anglicisms in German newspaper ads during that time period but starting at a higher level (15% to 44%). Two further studies compared the number of anglicisms in the *Dictionary of European Anglicisms* (Görlach, 2001) listed as being used in German, French, Italian or Spanish (Müller, 2003; Lohmann *et al.*, 2004). Both found that most anglicisms are used in German and numbers for French, Italian and Spanish are lower but non-negligible. While this evidence suggests that anglicisms are less frequently used in French than in German, their increasing usage in French is indisputable. In fact, Humbley (2006) infers from his and previous analyses that various European languages are all affected by English with French being no exception. In the French newspaper text published in *ZDNet France*[7] on IT-related subjects, the evaluation data used to assess the English inclusion classifier for French, the percentage of anglicisms amounts to 6.1% and 6.8% of tokens in the development and test set, respectively (see Table 4.1 in Chapter 4). These percentages are very similar to those found in the German evaluation data from the internet domain which was also annotated as part of this thesis project (6.0% of tokens in the development set and 6.4% in the test set listed in Table 3.1). These findings provide some evidence that the language policy advocated by the French government failed, at least in some sectors of French society. Nevertheless, it is difficult to arrive at clear conclusions regarding the intensity of the influence that English has on various languages in such cross-language comparisons.

2.1.3.3 Attitudes towards English Inclusions

Compared to Germany, France's government is playing a considerably more active role in protecting its language from foreign influences. However, it is unclear what the French people's attitude is to anglicisms given that they need to obey laws that restrict their usage of English terms and advocate the use of French translations instead. However, particularly French youngsters and people working in the media are relatively open to the use of anglicisms which can be attributed to the prestige of English in certain circles and the fact that French equivalents do not get introduced in time for tight deadlines in a dynamic and fast moving journalism environment. Therefore, a number of language purists, including several Francophone institutions, vehemently condemn the contamination of the French language with English forms (e.g. Laroch-

[7] `http://www.zdnet.fr/`

Claire, 2004). Whether this attitude is shared by the French people is arguable. Flaitz (1993), who acknowledges the fact that the use of English is widespread in France, reports the results of a survey on French people's attitudes towards English language and culture. His survey, participated in by 145 people living in Paris, Rouen, Troyes, and Montbard, shows that attitudes towards the English language were generally positive in 1993. Only 25% of subjects were worried about the influence of American English on the French culture. This led Flaitz to conclude that the position of the general French population on the use of anglicism diverges from that of the French power elite.

2.1.3.4 Types of English Inclusions

The types of English forms appearing in French, and recognised by the English inclusion classifier, are similar to those occurring in German (see Section 2.1.2.4). This section provides a brief overview of different types and examples, particularly those that are specific to French. As in German, French contains English borrowing (e.g. *browser*) and code-switching (e.g. *what you see is what you get*). Borrowings can be truncated and still refer to the English long forms like *foot* (football) or *snack* (snack bar) (Pergnier, 1989; Sokol, 2000). French also contains mixed-lingual forms, in particular English verbs with French endings (e.g. *coach*er for *to coach*) (Rollason, 2005). Such concatenated mixed-lingual forms are currently not recognised by the classifier.

A particular characteristic of anglicisms in French is the frequent usage of English present particle/gerund forms, the French "love affair" with *-ing* forms. They have generally taken on a locative meaning in French (Nicholls, 2003). For example, *bowling*, *camping*, *dancing*, *living*, and *parking* in French all refer to the locations where these activities take place. Sometimes English *-ing* forms appearing in French did not originally appear in English in their inflected form and with their particular meaning like *lifting* (face lift) (Thogmartin, 1984).

English compounds in French tend to be concatenated with or without hyphens, for example *junkmail* or *shake-hand* (handshake) (Tattersall, 2003; Thogmartin, 1984). The latter anglicism is also an example of order reversal. Some English compounds appearing in French do not exist in English and are therefore considered to be pseudo-anglicisms, e.g. *tennisman* (male tennis player) or *recordman* (male record holder). Pseudo-anglicisms are generally widespread in French, including examples such as *smoking* (tuxedo, dinner jacket), *zapping* (channel hopping), *lifting* (face-lift), or *spot*

(television commercial) (Picone, 1996; Nicholls, 2003).

2.1.4 English in Other Languages

The discussion above argued that occurrence of anglicisms is a common phenomenon in European languages. This has resulted in a plethora of research on this topic including the compilation of the *Dictionary of European Anglicisms* (Görlach, 2001) appearing in 16 different European languages and a series of studies presented, for example, at the international conference on *Anglicisms in Europe* held at the University of Regensburg in Germany in 2006 (Fischer , forthcoming). However, this language mixing phenomenon is not only limited to Europe. English is influencing many other languages with different types of alphabets. For example, English is frequently found in many Asian and Eastern European languages, including Chinese, Japanese and Korean as well as Hindi and Russian. The following section presents an account of anglicisms appearing in these languages. This is not an exhaustive summary of the types of English inclusions occurring in all of the world's languages. Instead, it presents evidence of anglicisms appearing in languages with non-Latin scripts and supports the argument that English is influencing different languages in various ways.

In Chinese, English loan words tend to appear as either Chinese translations or **transliterations**, i.e. transcriptions using Chinese characters, called **hanzi**, to approximate the English pronunciation. As each individual hanzi has one or more meanings, a transliteration of an anglicism can be meaningless to someone who fails to recognise it as foreign. For example, the English word *bus* is transliterated into Chinese using the two hanzi pronounced as *bā* (to hope, to wish) and *shì* (scholar, warrior, knight).[8] Such transliterations often result in an approximate pronunciation of the original sounds. Pronunciation variation in different dialects can even lead to multiple transliterations for one anglicism. Although translations and transliterations tend to be preferred, English is also appearing written in its original Latin characters in some sectors of Chinese society (Hall-Lew, 2002). Often poor translation from Chinese to English, also referred to as **Chinglish**, results in the use of terms that seem bizarre or even humorous to a native English speaker like *No noising* (meaning *No shouting*) or *Don't Bother* (meaning *Do not disturb*).[9] Such mistranslations are occurring to such

[8]Example taken from: `http://www.yellowbridge.com/language/englishloan.html`

[9]Examples taken from: http://news.bbc.co.uk/1/hi/world/asia-pacific/6052800.stm

an extent in Chinese that Bejing's municipal government has decided to stamp them out in the run-up to the Olympics in 2008 (BBC News, 15th of October 2006)[10].

Japanese writing consists of a combination of three types of scripts: **kanji** (Chinese characters) as well as **hiragana** and **katakana** (two phonetic, syllabic scripts). Foreign words and proper names are generally transliterated using katakana characters which do not carry a meaning by themselves in the same way as kanji (or hanzi) do. Japanese words are generally written either in kanji or hiragana characters which makes words of foreign origin relatively straight-forward to distinguish (Breen, 2005). However, katakana are also used for certain technical and scientific expressions, for onomatopoeia (words imitating sounds) or for the purpose of emphasising words in text (similarly to the italic font used in Latin-based scripts). As in Chinese, the pronunciation of foreign words is adapted by Japanese speakers to their native tongue and therefore differs from that of English native speakers (e.g. see Shirai, 1999, on the germination in loans from English to Japanese). The main reason is that some English characters and sounds do not exist in Japanese.

Yoneoka (2005) found a large overlap of the English vocabulary used in Japanese and Korean. In fact, some English borrowings enter Korean via Japanese and therefore carry some Japanese phonological characteristics (Yoneoka, 2005). For example the English expression *ok*, *okebari* in Korean, is heavily influenced by the Japanese word *okimari* (to decide). English terms are used frequently in Korean, either in their original forms or in **Hangul** transliterations or translations (Bae and L'Homme, 2006). Hangul is the official Korean script which is a phonetic, syllabic alphabet whose individual characters do not convey meaning. The recent influx of anglicisms both into Korean and Japanese is largely attributed to globalisation and the growing spread of computer technology. Foreign names of new technologies are therefore often directly inserted without transliteration. This results in language mixing, also referred to as **Janglish** and **Konglish** (or wasei eigo), which can leave English native speakers bewildered. The Latin alphabet is also used for foreign abbreviations and acronyms or for English words in lyrics of Japanese popular music (Moody, 2006).

The genre of popular culture across Asia is renowned for mixing native languages with English. However, motivations behind this language mixing can vary from country to country. Unlike in Chinese, Japanese and Korean where the use of English

[10]Article published at: `http://news.bbc.co.uk/1/hi/world/asia-pacific/6052800.stm`

is seen as a sign of asserting one's identity or showing resistance to traditions and customs, Kachru (2006) states that the use of English in Indian pop culture is more of a playful manner. Kachru examined **Hinglish**, a mix of English and Hindi, in popular songs from Bollywood movies. He concludes that English has been integrated to such an extent in Hindi that it is no longer perceived as an unfamiliar language. English words are either borrowed directly or invented. In advertisements, entire English slogans are used, like *We need your heads to run our business* on the front of a barber's shop in Juhu, Mumbai (Kachru, 2006). The influx of English into Hindi TV commercials, news articles and film song lyrics has increased according to Kachru. Language mixing occurs on the sentence, word and sub-word level as exemplified in the song shown in Figure 2.5 where English words are marked in italic font. However, Hinglish was not always as popular as it currently is. For example, Sanjay Sipahimalani, Executive Creative Director of Publicis India, said that ten years ago Hinglish would have signalled a lack of education but today it is a huge asset for his agency (cited in Baldauf, 2004). This shows that the motivation behind using anglicisms can change over time.

In Russian, most English words and expressions are embedded as transliterations in the Cyrillic alphabet, e.g. ваучер (voucher, specifically privatisation vouchers). Anglicisms can also end in Russian inflections like ваучеризациа (voucherisation), referring to the issuing of privatisation vouchers, an expression that is not commonly used in English (Dunn, 2007). Moreover, pseudo-anglicisms like инвентор (inventor, referring to an events organiser in Russian) are very common in Russian. Ustinova (2006) speaks of an invasion of English words in Russian that raises significant concern to Russian legislative and executive authorities that want to take legislative measures against such language mixing. This increase in the use of English is supported by Ustinova's findings when examining the frequency of English in Russian advertisements. She found that 76% of Russian TV commercials contain English or a mixture of English and Russian. The main function of English expressions in advertisements is to express novelty and prestige and signal high quality products. For this purpose, some English names are not always transliterated into the Cyrillic alphabet but advertised in their original Latin script. This is also the case for some English expressions like *dress code* or *face control* (also used in Cyrillic script: фейс-контроль) referring to the job

Excuse Me... Kyaa Re	Excuse me...What's it?
Meraa Dil Tere Pe Fidaa Re...	I am smitten by you
Bus Stop Pe Dekhaa Tujhe Pehli Baar	I saw you first at the bus stop
JhaTke MeN Ho Gayaa Tere Se Pyaar	I fell in love as the bus lurched
Excuse Me... aaN Bolnaa	Excuse me...yes?
MaiN Pehle Se Shaadi Shudaa Re...	I am already married.
Abhi To HooN Saalaa *Road*pati	I am the master of street
Ladki ChaahuuN KaroRpati...	I want a billionaire girl
Race Course MeN Dekha Tujhe Pehli Baar	I saw you at the Race Course
Counter Pe Ho Gayaa Tere Se Pyaar	I fell in love at the counter
Excuse Me...Kyaa Re	Excuse me...What's it?
Police MeN Hai Meraa MiyaaN Re	My husband is in the police
...*Excuse Me...Yes Please*	Excuse me...Yes please
Ban Jaa Mera Bhaiyaa Re	Better become my brother

Figure 2.5: Hinglish song lyrics, example taken from Kachru (2006)

of a bouncer outside a club when deciding who can enter or not (Dunn, forthcoming).

It can be concluded that English is influencing many languages. The anglicisation of those languages is a complex process with various reasons and motivations. Foreign words or proper names pose substantial difficulties to NLP applications, not only because they are hard to process but also because, theoretically, they are infinite in number. Moreover, it is impossible to predict which foreign words will enter a language, let alone to create an exhaustive gazetteer of them. Therefore, the increasing use of English forms in different languages presents a challenge to NLP systems that are conventionally designed to handle monolingual text. The task of recognising English inclusions is the main focus of the work presented in this thesis. Before introducing an automatic classifier able to detect English language portions in otherwise monolingual text (Chapters 3 and 4), this chapter will review previous work on automatic, specifically mixed-lingual, language identification (Section 2.2).

2.2 Automatic Language Identification

Written language identification (LID) is the task of automatically determining the language of an electronic document. This is an important issue in many areas of multilingual information processing such as indexing or text classification. In our increasingly multilingual society, automatic LID is vital to enable accurate processing of natural language in the case where a document's language is not apparent for example from its metadata, or from background information. Whilst first research initiatives in written LID began 30 years ago, the 1990s saw increasing efforts in this field as the first multilingual corpora were published. Today, a variety of off-the-shelf LID systems are available. Presented with several lines of text, they will attempt to identify its language. This chapter will first give a brief overview of conventional approaches to LID and then report specifically on LID systems designed to deal with mixed-lingual text.

The majority of existing state-of-the-art LID systems rely on word-level information such as diacritics and special characters (Newman, 1987), common short words (Johnson, 1993), character-level language models (Dunning, 1994) or methods based on character-based n-gram frequencies (Beesley, 1988; Cavnar and Trenkle, 1994; Damashek, 1995; Ahmed *et al.*, 2004). Other automatic LID programs function by determining the character encoding of a document (Kikui, 1996). Comparisons of different techniques (Grefenstette, 1995; Capstik *et al.*, 1999; Padró and Padró, 2004; Kranig, 2005) demonstrate that it is difficult to determine the one best LID method.[11] Their results largely depend on the type and number of languages involved, the amount of training data and the number of input words. This means that LID accuracy increases with the length of the test data and is not satisfactory for individual words.

Cavnar and Trenkle (1994)'s approach of using character n-gram frequency lists to determine the language of a new piece of text is the underlying algorithm of TextCat, an automatic LID system developed by Gertjan van Noord.[12]. It is used for comparison in some of the experiments described in Chapter 3. It creates n-gram frequency lists (1 to 5-grams) for various language corpora and the text to be identified (n-gram profiles) which are sorted by frequency counts in ascending order. The n-gram profile of the text is then compared to each of the language corpus profiles in terms of n-gram ranks

[11] For a summary of different LID methods see also Cole *et al.* (1997).

[12] `http://www.let.rug.nl/~vannoord/TextCat`

by calculating the sum of all absolute rank distances for each n-gram t_i. The language l resulting in the smallest distance D signals the language of the test document.

$$D = \sum_{i=1}^{N} |rank(t_i, text) - rank(t_i, l)| \tag{2.1}$$

TextCat provides corpus profiles for 69 languages each containing 400 n-grams per language. If an n-gram does not occur in the profile, it is given a maximum distance score of 400. Cavnar and Trenkle (1994) report an accuracy of 99.8% for input strings of at least 300 bytes and an accuracy of 98.3% for strings of less that 300 bytes using their LID approach. No information is given regarding its performance for really small strings. In the experiments presented in Sections 3.4.1 and 3.4.3.1, the collection of corpus profiles was limited to the required languages, namely German and English.

Most automatic LID systems are successful in identifying the base language of a document but are not designed to deal with mixed-lingual text to identify the origin of individual foreign words and proper names within a given sentence. Initial work on this issue has been carried out by Font-Llitjós and Black (2001) who use Cavnar's n-gram statistics approach to estimate the origin of unseen proper names as a means of improving the pronunciation accuracy for TTS synthesis. While the LID performance (for 26 languages) is not evaluated separately, the pronunciation accuracy of proper names increases by 7.6% from the baseline of 54.08% when adding language origin probabilities as a feature to the CART-based decision tree model. LID has also been applied in the field of name transliteration. Qu and Grefenstette (2004) used LID as a way of determining language-specific transliteration methods for Japanese, Chinese and English named entities (NEs) written in Latin script. Their language identifier is based on tri-gram frequencies, whereby the language of a word is that for which the sum of the normalised tri-grams is highest. On the training data, LID yields accuracies of 92% for Japanese, 87% for Chinese and 70% for English names. A further study is that of Lewis *et al.* (2004) who implemented a character n-gram classifier to differentiate between the 10,000 most common English words, 3,143 unique transliterated Arabic and 20,577 Russian names. Each of the three data sets is divided into 80% training and 20% test data, a process which is repeated 4 times resulting in overlapping training and test sets. Average precision values amount to 81.1% for Russian names,

92.0% for Arabic names and 98.9% for English common words. When combining LID and language-specific letter-to-sound rule decision trees, the precision of phones in the system transcription that match hand-transcribed phones amounts to 89.2% compared to a baseline precision of 80.1% for a system that is simply trained on the CMU lexicon, a pronunciation lexicon for American English.

The latter work largely depends on the distinct statistical characteristics between languages. English, Russian and Arabic are very different languages. Although English and Russian are both Indo-European languages, they belong to different language groups, namely Germanic and Slavic, respectively. Arabic, on the other hand, is a member of the Afro-Asiatic language family. Therefore, LID for NEs of more closely related languages is anticipated to be a more challenging task. Another interesting point made by Lewis *et al.* (2004) is that unseen foreign words in English documents are generally proper names. While this is largely true for English, the same cannot be said for German text, for example, in which increasing numbers of anglicisms have been recorded, particularly in the last 50 years. This can be mainly attributed to technological advances, in particular the invention of the computer and the internet, as well as political events such as the creation and enlargement of the EU. As a result, German documents frequently contain English inclusions, not only NEs but also many other content words. The influx of anglicisms into German and other languages was examined in detail in Sections 2.1.2 to 2.1.4.

It is evident that LID information would be beneficial to multilingual TTS synthesis and other NLP applications that need to handle foreign names. However, with the increasing influence that English has on other languages, state-of-the-art systems must also be able to deal with other types of foreign inclusions such as English computer terminology, expressions from the business world or advertising slogans that are encountered in texts written in other languages. Moreover, such language mixing does not only happen on the word level, i.e. a German sentence containing some foreign words. It also occurs on the morpheme level when a word contains morphemes from different languages. First efforts that address mixed-lingual LID are discussed in the following section.

2.2.1 Language Identification of Mixed-lingual Data

Most conventional LID systems are successful in recognising the language of larger portions of text but are not well suited to classify individual tokens or sub-parts thereof. This section examines four LID approaches that are designed to deal with mixed-lingual input text. The first method relies on **morpho-syntactic analysis** combined with **lexicon lookup** (Pfister and Romsdorfer, 2003). The second approach is built on a combination of different methods including **dictionary lookup** and **character n-gram statistics** (Marcadet *et al.*, 2005). The third system combines **Hidden Markov Model** language tagging with dictionary lookup and character-based n-gram modelling (Farrugia, 2005). Finally, the last algorithm (Andersen, 2005) is based on combined **chargram and regular expression matching**. Each study is reviewed in detail.

2.2.1.1 Morpho-syntactic Analysis of Mixed-lingual Data: Pfister and Romsdorfer (2003)

Pfister and Romsdorfer (2003) outline the language mixing phenomena which is typically encountered in German texts and derive a method for analysing such data. Based on their analysis, they conclude that Swiss newspaper articles contain many foreign inclusions, the majority of which are of English but also some of French origin. These results are consistent with findings by Henrich (1988) who states that most of the foreign inclusions in German text for which German pronunciation patterns do not apply are either English or French. Such inclusions can vary from simple word stems to entire phrases. Pfister and Romsdorfer (2003) identified various types of English inclusions and grouped them into three major categories:

1. Mixed-lingual word forms produced from an English stem by means of German declension, conjugation or compounding, e.g.

   - `Den` *`Managern`* `wird misstraut.` (noun stem + *n*, dative plural case)
     Translation: *Managers are mistrusted.*

   - `Er` *`surft`* `gern.` (verb stem + *t*, 3rd person singular)
     Translation: *He likes to surf.*

- Das ist ein *smart*es System. (adjective stem + *es*, nominative neuter case)
  Translation: *This is a smart system.*

- *Manager*gehälter sind umstritten. (compound noun)
  Translation: *Manager salaries are controversial.*

- Exotic mixed-lingual words like: *outgesourc*t (outsourced).

2. Full foreign word forms that follow the foreign morphology, e.g.

   - Die *Fans* lieben ihr Team. (noun)
     Translation: *The fans love their team.*

   - Der *Laser* ist eine Lichtquelle. (noun)
     Translation: *The laser is a light source.*

   - Sie ist *happy*. (adjective)
     Translation: *She is happy.*

3. Multi-word inclusions which are syntactically correct foreign constituents, e.g.

   - Der Konkurs von *Swiss Dairy Food* ist ... (proper name)
     Translation: *The bankruptcy of Swiss Dairy Food is ...*

   - *Human Touch* kommt an. (noun group)
     Translation: *Human touch goes down well.*

By foreign language inclusions they refer to foreign words which are less assimilated into the receiver language and tend to keep their foreign pronunciation. Pfister and Romsdorfer (2003) make a distinction between foreign inclusions and assimilated loan words which are more integrated into the base language in terms of morphology, syntax and pronunciation. A system able to derive the appropriate pronunciation and prosody for mixed-lingual text can be used for polyglot TTS synthesis. While such a system can deal with assimilated loan words as with other words in the base language

of the input, it must analyse which sections of the input are foreign and do not follow base language patterns. In this case, one system analyses the input and produces synthesised output with one voice. Therefore, polyglot TTS differs from multilingual TTS where independent subsystems are applied for input in different languages and output is synthesised in the respective language-specific voices (see Section 6.1.4).

Due to the language mixing phenomenon on the sentence and word level in German, Pfister and Romsdorfer (2003) conclude that morphological and syntactic analysis is required to process foreign inclusions for TTS synthesis. They argue that the use of a lexicon containing full word forms is not sufficient as the number of mixed-lingual words is large, particularly due to the virtually arbitrary combinations of stems, endings and prefixes in the case of verbs and the unlimited number of mixed-lingual compounds. There also is the issue of homographs belonging to different languages which occur more frequently in mixed-lingual documents. The example cited is the word *argument* which could be German, English or French depending on the context.

Pfister and Romsdorfer (2003) then describe an approach of mixed-lingual text analysis for a set of languages $\{L_1, L_2, L_3, \ldots\}$. First, a set of monolingual analysers is designed comprising language-specific lexica and word and sentence grammars. Secondly, an inclusion grammar is established for each language pair $\{L_i, L_j\}$ which defines the elements of language L_j that are allowed as foreign inclusions in language L_i. A mixed-lingual morpho-syntactic analyser for German and English (German being the base language) would require the loading of the lexica and grammars of both languages as well as the inclusion grammar G_{GE} (Figure 2.6). This approach is similar to that proposed by Joshi (1982) for dealing with Marathi-English code-switching where an asymmetric switching rule allows for the exchange of categories in the base (or matrix) language grammar with those in the embedded language grammar.

The language-specific grammars are independent and grammar rule penalty values are used to determine the optimal solution. The penalty values are set by linguistic experts with the aim of solving interlingual ambiguities. All ambiguities in the chart of a full parse are kept and the final sentence is selected according to the minimum accumulated penalty points. The morpho-syntactic analyser marks each morpheme with the corresponding language identifier and pronunciation. Figure 2.7 presents the analyser's output for the German sentence "*Er surft im World Wide Web.*".

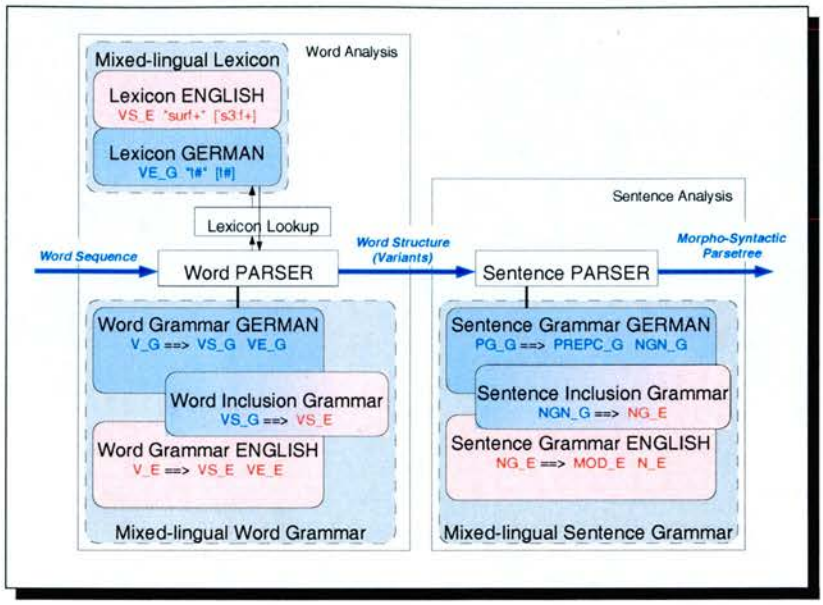


Figure 2.6: Mixed-lingual German/English analyser (Romsdorfer and Pfister, 2003)

Pfister and Romsdorfer (2003) claim that the use of the inclusion grammar provides an appropriate solution to the problem of interlingual homographs. The rules of the inclusion grammars allow all variants of a homograph in different languages, but the high penalty value prioritises the one that matches the language of the including constituent. When having to distinguish between the languages English, French and German for the word *argument*, the inclusion grammar will give a higher penalty to the foreign language variant and consequently prioritise the base language. This approach works for the majority of cognates which are more likely to occur in the base language of the text. However, it does not present a viable solution for determining the language of interlingual homographs that are actually foreign inclusions possibly with the same gender but not necessarily the same semantics as the base language variant. Consider the following sentence:

(2) Ich habe nur ein **Lager** getrunken.

Translation: *I only drank one lager.* OR *I only drank one camp.*

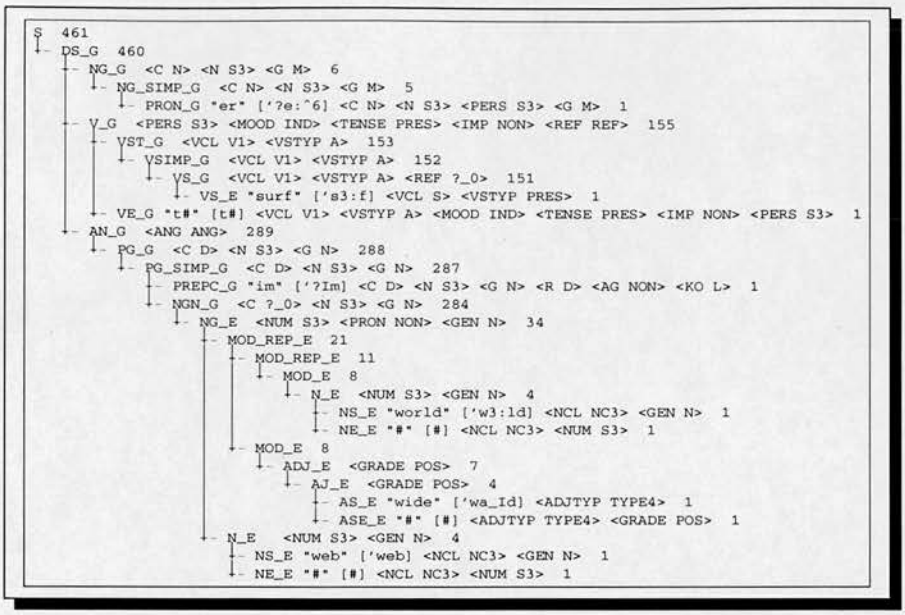


Figure 2.7: Mixed-lingual analyser output (Pfister and Romsdorfer, 2003)

The interlingual homograph *Lager* could either refer to the German neuter noun, which has many different meanings including *camp*, or the English noun, i.e. a type of beer. Given the context, it is evident that we are dealing with an English inclusion. However, the analyser would give higher priority to the German variant.

Pfister and Romsdorfer (2003) note that individual language grammars consist of more than 500 rules whereas the inclusion grammar contains around 20 rules. While they state that the morpho-syntactic analyser is precise at detecting the language of tokens and the sentence structure, they did not actually evaluate the performance of their system. The reason is that the various lexica and grammars used by the rule-based morpho-syntactic analyser are relatively small and coverage is thus very limited (correspondence with authors). Results would be largely dominated by the words that are not covered by the morphological analyser and not measure the performance of the approach. It is therefore unclear how well the analyser performs on real mixed-lingual data. Although Pfister and Romsdorfer (2003) have taken an interesting approach to dealing with mixed-lingual data, a system working with larger lexica and grammars will be very expensive in terms of computational overhead and may fail when rules contradict each other. This method is also very costly considering that linguistic experts have to write from scratch all the necessary grammars for each language scenario.

2.2.1.2 Combined Dictionary and N-gram Language Identification: Marcadet *et al.* (2005)

A further approach to LID for mixed-lingual text is that of Marcadet *et al.* (2005). Their LID system is specifically designed to function at the front-end to a polyglot TTS synthesis system. They present experiments with a dictionary-based transformation-based learning (TBL) and a corpus-based n-gram approach and show that a combination of both methods yields best results.

The dictionary TBL approach is based on the concept of starting with a simple algorithm and iteratively applying transformations to improve the current solution. It starts with an initial state annotator which classifies tokens as either English, French, German, Italian or Spanish based on dictionary lookup. The dictionary contains the most frequent words for each language and is severely reduced in size by applying over 27,000 morphological rules including special character as well as suffix and prefix rules. Marcadet *et al.* (2005) do not give any details as to how these rules are created. After the initial lookup, all tokens which could not be assigned to one specific language are treated as ambiguous. Subsequently, the primary language of each sentence is determined. Finally, the language ambiguous tokens are resolved by means of a rule tagger. This tagger is made up of 500 hand-written rules conditioning on the current, previous or next word and language tag. Even though the authors call this method their TBL approach, TBL is not actually carried out due to the lack of bilingual training data.

Their second method, the n-gram with context approach, is entirely corpus-based. A character n-gram language model is trained for each language and during the LID stage, the most likely language tag L for a word w is computed as:

$$\hat{L} = argmax_L\{P(L|w)\} \tag{2.2}$$

The language likelihood of a given word is calculated on the basis of the probability of its character n-gram sequence (7-grams) and weighted language likelihood scores of the previous and next token in order to account for context.

Marcadet *et al.* (2005) evaluate their system using three small mixed-lingual test scripts in different languages (Table 2.2). The proportion of foreign inclusions in each of the test scripts suggests that they are not a random selection of text but rather a col-

Test Data	French	German	Spanish
Sentences	50	49	25
English words	195	123	119
French words	1129	0	0
German words	6	795	2
Italian words	5	0	0
Spanish words	8	0	494
Punctuation	202	132	99
Total	1545	1050	714

Table 2.2: Language origins of words in three test scripts (Marcadet *et al.*, 2005)

lection of sentences specifically chosen for this task. The n-gram approach is outperformed by the dictionary TBL approach. Combining both methods and subsequently running the rule tagger outperforms both individual scores. The dictionary lookup effectively deals with known tokens and the n-gram method resolves unknown words. The combined approach yields word error rates (WERs) of 0.78 on the French data, 1.33 on the German data and 0.84 on the Spanish data, respectively (Table 2.3).

	n-gram	TBL	combined	gain
French	6.73	1.36	0.78	42.65
German	2.86	2.67	1.33	50.19
Spanish	4.90	1.40	0.84	40.00

Table 2.3: Token-based language identification error rates (in percent) for three different test scripts and methods (Marcadet *et al.*, 2005)

Table 2.2 shows that the test sets are very small. Moreover, the test sentences are not randomly selected. It would therefore be interesting to determine the system's performance on random data and how it scales up to larger data sets. The paper also does not report the performance of each individual approach (initial state annotator, n-gram and context rules) in the combined system or the performance for each language

separately in terms of precision and recall. Their test sets only contain a very small number of non-English foreign inclusions. As individual language scores are not given, it is unclear if their system identifies them correctly. It may be sufficient to concentrate on English inclusions alone to make the system more computationally attractive.

Section 3.4 examines how the English inclusion classifier developed as part of this thesis performs on the German test data designed by Marcadet *et al.* (2005) and shows that it marginally outperforms their system. The English inclusion classifier therefore compares favourably to state-of-the-art mixed-lingual LID.

2.2.1.3 Hidden Markov Model Language Tagging: Farrugia (2005)

Similar to the previous approach, the following method is also designed to function as a pre-processing step for TTS. Farrugia (2005) proposes token-based LID for mixed-lingual Maltese English SMS messages by means of Hidden Markov Model (HMM) language tagging combined with dictionary lookup and a character n-gram language model for dealing with unknown words.

In the HMM, the language tags of token are the hidden states (x) and the words are the observations (y). Figure 2.8 shows that the current language tag (x_t) is dependent on the previous language tag (x_{t-1}), and the currently observed word (y_t) is dependent on the current language tag (x_t).

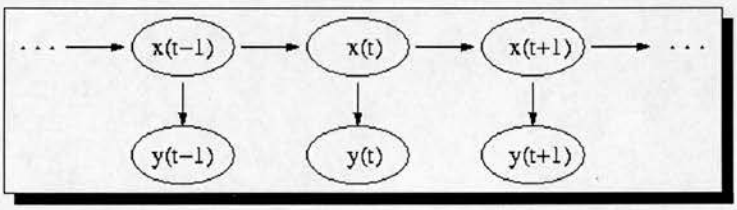


Figure 2.8: Hidden Markov Model architecture, source: `http://en.wikipedia.org/wiki/Hidden_Markov_model`

These dependencies are captured via transition probabilities between states (α) and emission probabilities between states and observations (β). α and β as well as the initial state distribution (π) are all computed from an annotated SMS training corpus. Given these parameters, the aim is to determine the most likely sequence of language tags

that could have generated the observed token sequence. This is done by means of the Viterbi algorithm (e.g. Rabiner, 1989).

The LID algorithm handles unknown words first by means of a dictionary lookup for each language involved. If an unknown token is present in a dictionary, four training samples are added with the corresponding language tag. If it is not found in the dictionary, one training sample is added. If a token is not found in any dictionary, the system backs off to a character n-gram language model based on a training corpus for each language (e.g. Dunning, 1994). Farrugia uses a parallel Maltese English corpus of legislative documents for this purpose. Three samples are then added to the SMS training corpus for the most likely language guess. After biasing the training sample in this way, the HMM is rebuilt and the input text is tagged with language tags.

Farrugia's algorithm is set up to distinguish between Maltese and English tokens. He reports an average LID accuracy of 95% for all tokens in three different test sets containing 100 random SMS messages each, obtained via a three-fold cross-validation experiment. As the language distribution for each of the test sets is not provided, it is unclear how well the system performs for each language in terms of precision, recall and F-score and consequently how proficient it is at determining English inclusions. Therefore, it is difficult to say what improvement this LID system provides over simply assuming that the input text is monolingual Maltese or English. In fact, Farrugia (2005) does not clarify at what level code-switching takes place, i.e. if SMS messages are made up of mostly Maltese text containing embedded English expressions, if language changes are on the sentence level, or if messages are written entirely in Maltese or English. Furthermore, it would be really interesting to investigate how well Farrugia (2005)'s approach performs on running text in other domains and what the performance contribution of each of the system components is. Considering that languages are constantly evolving and new words enter the vocabulary every day, the dictionary and character n-gram based approach for dealing with unknown words is relatively static and may not perform well for languages that are closely related.

2.2.1.4 Lexicon Lookup, Chargrams and Regular Expression Matching: Andersen (2005)

Andersen (2005) notes the importance of recognising anglicisms to lexicographers. He tests several algorithms based on lexicon lookup, character n-grams and regular

expression matching and a combination thereof to automatically extract anglicisms in Norwegian text. The test set, a random sub-set of 10,000 tokens from a neologism archive (Wangensteen, 2002), was manually annotated by the author for anglicisms. For this binary classification, anglicisms were defined as either English words or compounds containing at least one element of English origin. Based on this annotation, the test data contained 563 tokens classified as anglicisms.

Using lexicon lookup only, Andersen determines that exact matching against a lexicon undergenerates in detecting anglicisms, resulting in low recall (6.75%). Conversely, fuzzy matching overgenerates, resulting in low precision (8.39%). The character n-gram matching is based on a chargram list of 1,074 items constisting of 4-6 characters which frequently occur in the British National Corpus (BNC). Being typical English letter sequences, any word in the test set containing such a chargram is classified as English. This method leads to a higher precision of 74.73% but a relatively low recall of 36.23%. Finally, regular expression matching based on English orthographic patterns results in a precision of 60.6% and a recall of 39.0%.

On the 10,000 word test set of the neologism archive (Wangensteen, 2002), the best method of combining character n-gram and regular expression matching yields an accuracy of 96.32%. Simply assuming that the data does not contain any anglicisms yields an accuracy of 94.47%. Andersen's reported accuracy score is therefore misleadingly high. In fact, the best F-score, which is calculated based on the number of recognised and target anglicisms only, amounts to only 59.4 (P = 75.8%, R = 48.8%). However, this result is unsurprisingly low as no differentiation is made between full-word anglicisms and tokens with mixed-lingual morphemes in the gold standard.

A shortcoming of Andersen's work, and other reviewed studies, is that the methods are not evaluated on unseen test data. The knowledge of previous evaluations could have affected the design of later algorithms. This could easily be tested on another set of data that was not used during the development stage. It would also be interesting to investigate how the methods devised by Andersen perform on running text instead of a collection of neologisms extracted from text. While Andersen's work is already applied in a language identification module as part of a classification tool for neologisms, language identification on running text could exploit knowledge of the surrounding text. Applied in such a way, anglicism detection would also allow lexicographers to examine the use of borrowings in context.

2.3 Chapter Summary

This chapter presented an overview of the entire background and theory behind the work presented in this thesis. It first discussed the issue of language mixing with English as a result of globalisation and the omnipresence of the internet. It was found that English influences many languages and that the influx of anglicisms is on the increase. Many different types of language mixing phenomena and different motivations for using English were established in the analysis of different languages. Particular focus was given to German and French.

As many NLP applications are relying on monolingual text input, the issue of anglicisation of languages needs to be addressed in order to improve the accuracy of such systems. This task of recognising foreign words in different languages is starting to be addressed. Previous studies reviewed in this thesis rely on lexicon lookup, character n-gram statistics or rule-based morpho-syntactic analysis in order to detect foreign inclusions. However, none of the proposed methods were evaluated on unseen data or running text. Moreover, some of the methods proposed require training data or rules which linguists need to design from scratch for every new language scenario.

In the following three chapters, the English inclusion classifier, designed as part of this thesis project, is introduced and evaluated in detail. The complete classifier and its components are evaluated intrinsically on German and French data sets (Chapters 3 and 4) and extrinsically in several parsing experiments. All evaluation metrics and notations are presented in Appendix A in order to facilitate better understanding of all experiments described in the remainder of this thesis.

Chapter 3

Tracking English Inclusions in German

The recognition of foreign words and foreign named entities (NEs) in otherwise monolingual text is beyond the capability of many existing LID approaches and is only starting to be addressed. This language mixing phenomenon is prevalent in German where the number of anglicisms has increased considerably in recent years. This chapter presents an annotation-free and highly efficient system that exploits linguistic knowledge resources, namely English and German lexical databases and the World Wide Web, to identify English inclusions in German text (Alex and Grover, 2004; Alex, 2005). This system is referred to as the English inclusion classifier.

After briefly reiterating the issue of English inclusions and motivating the tool in Section 3.1, Section 3.2 describes the corpus which was collected and annotated specifically for this task as well as some annotation issues that arose in the process. This chapter then continues with a detailed overview of the system modules of the English inclusion classifier (Section 3.3). The final system as well as individual components are evaluated in Section 3.4. Additionally, the performance of the classifier on unseen test data is presented and compared to another state-of-the-art mixed-lingual LID approach. The final design of the English inclusion classifier is based on the results of a series of parameter tuning experiments which are presented in Section 3.5. Finally, the system's performance is compared to the performance of a supervised machine learner in a series of in- and cross-domain experiments with a maximum entropy (maxent) tagger trained on a hand-annotated corpus (Section 3.6).

3.1 Motivation

In natural language, new inclusions typically fall into two major categories, foreign words and proper nouns. They cause substantial problems for NLP applications because they are hard to process and frequent in number. It is difficult to predict which foreign words will enter a language, let alone create an exhaustive gazetteer of them. In German, there is frequent exposure to documents containing English expressions in business, science and technology, advertising and other sectors. The increasing influence which English is having on German is also referred to as Denglish (German mixed with English) and widely discussed in the German media. Having a look at newspaper headlines confirms the existence of this phenomenon (Weiss, 2005):

(1) **Security-Tool** verhindert, dass **Hacker** über **Google** Sicherheitslücken finden.

Translation: *Security Tool prevents hackers from finding security holes via Google.*

Foreign word inclusions can be regarded as borrowings which are further sub-divided into *assimilated loan words* and *foreign words*. Loan words are relatively integrated into the receiver language whereas foreign words are less integrated (Yang, 1990). The system described here is specifically tailored to recognise foreign words and names with English origin. However, the system also attempts to identify words with the same spelling in both languages, including assimilated loan words and internationalisms stemming from English and other languages.[1]

The benefit which the automatic classification of English inclusions presents to natural language parsing will be determined in a task-based evaluation in Chapter 5. As foreign inclusions carry critical content in terms of pronunciation and semantics, their correct recognition will also provide vital knowledge to many applications that process natural language, including polyglot text-to-speech synthesis and machine translation.

[1]Loan substitutions (Lehnprägungen, Betz (1974)) or internal borrowing (inneres Lehngut, Yang (1990)), like the word *Spracherkennung* (speech recognition), are other types of borrowings. These are instances where the lexical items of the donor language are expressed using semantically identical or similar lexical items of the receiver language. For the purpose of the experiments, loan substitutions are not separately identified and are classified as German words as they are made up of German morphemes.

These applications are elaborated on in more detail in Chapter 6. In the same chapter, the English inclusion classifier will also be presented as a valuable tool for linguists and lexicographers who study this language-mixing phenomenon as lexical resources need to be updated and reflect this trend.

3.2 Corpus Description and Preparation

3.2.1 Data

As the classification of foreign inclusions is a relatively novel computational linguistics task, there was no appropriate hand-annotated data set available at the outset of this research project. This lead to the collection of a development and test corpus made up of a random selection of German newspaper articles from the Frankfurter Allgemeine Zeitung.[2] The articles were published between 2001 and 2005 in the domains: (1) **internet & telecoms**, (2) **space travel** and (3) **EU**. These specific domains were chosen to examine the use and frequency of English inclusions in German text of a more technological, political or scientific nature. The decision to randomly select was a deliberate one, as one of the aims was to determine the typical frequency of English inclusions readers can expect in texts written in those three domains. With approximately 16,000 tokens per domain, the overall development corpus comprises 48,000 tokens (see Table 3.1) in total. The test set is of approximately equal size as the development set for each domain. It was ensured that the articles in the test data do not overlap with those in the development data. The test set was treated as unseen and only used to evaluate the performance of the final system.

3.2.2 Annotation

In order to evaluate the performance of the English inclusion classifier quantitatively, an annotated gold standard was required. The initial classifier[3] output was used as a basis for hand annotation, i.e. the output was loaded into the annotation tool and corrected manually. The gold standard annotation was conducted using an annotation

[2] `http://www.faz.net`

[3] The initial classifier was a combination of the lexicon and search engine modules described in Section 3.3.3 and 3.3.4 but without post-processing or fine-tuning.

tool based on NXT (Carletta *et al.*, 2003) which operates with stand-off XML input and output. The binary annotation distinguishes between two classes using the BIO-encoding (Ramshaw and Marcus, 1995): English inclusion tokens are marked as I-EN (inside an English token) and any token that falls outside this category is marked as O (Outside). As the annotation was performed on the level of the token (and not phrase), an English inclusion received the tag B-EN only if it was preceded by another English inclusion. The annotation guidelines, which are presented in detail in Appendix B, specified to the annotators to mark as English inclusions:

- all tokens that are English words even if part of NEs: *Google*
- all abbreviations that expand to English terms: *ISS*
- compounds that are made up of two English words: *Bluetooth*

For the evaluation, it was also decided to ignore English-like person and location names as well as English inclusions occurring:

- as part of URLs: *www.stepstone.de*
- in mixed-lingual unhyphenated compounds: *Shuttleflug* (shuttle flight)
- with German inflections: *Receivern* (with German dative plural case inflection)

Further morphological analysis is required to recognise these. These issues will be addressed in future work when mixed-lingual compounds and inflected inclusions also need to be represented in the gold standard annotation.

Table 3.1 provides some corpus statistics for each domain and presents the number of English inclusions annotated in the various gold standard development and test sets. Interestingly, the percentage of English inclusions varies considerably across all three domains. There are considerably more English tokens present in the articles on the internet & telecoms and space travel than in those on the EU. 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 official languages, currently numbering 23. This language policy renders English less dominant in this domain than was

Data		Development Set					Test Set				
Domain		Tokens	%	Types	%	TTR	Tokens	%	Types	%	TTR
IT	Total	15919		4152		0.26	16219		4404		0.27
	English	963	6.0	283	6.8	0.29	1034	6.4	258	5.9	0.25
SP	Total	16066		3938		0.25	16171		4315		0.27
	English	485	3.0	73	1.9	0.15	456	2.8	151	3.5	0.33
EU	Total	16028		4048		0.25	16296		4128		0.25
	English	49	0.3	30	0.7	0.61	173	1.1	86	2.1	0.50

Table 3.1: English token and type statistics and type-token-ratios (TTR) in the German development and test data sets.

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.

While the proportion of English inclusions is relatively similar both in the development and test sets on internet & telecoms (6.0 versus 6.4%) and space travel (3.0 versus 2.8%), the test set on the EU contains considerably more English inclusions (1.1) than the EU development set (0.3). Regarding the development data, the type-token ratios (TTRs) signal that the English inclusions in the space travel data are least diverse (0.15). However, in the test data, the internet-related articles contain the most repetitive English inclusions (0.25). Even though the articles are a random selection, it is difficult to draw definite conclusions from these numbers as the data sets are small.

Table 3.2 lists the five most frequent English inclusions in each development set, covering various types of anglicisms that have entered the German language. All examples demonstrate the increasing influence that English has on German. First, there are English terms such as *Internet* whose German equivalents, in this case *Netz*, are rarely used in comparison. This is reflected in their low frequency in the corpus. For example, *Netz* only appeared 25 times in all of the 25 IT articles in the development set, whereas *Internet* appeared 106 times in the same set of articles. The German term was only used 19% of the time. This result corresponds to the findings by Corr (2003) which show that Germans tend to favour the use of anglicisms referring to specific computer vocabulary over that of their German translations.

Internet		Space		EU	
Token	f	Token	f	Token	f
Internet	106	ISS	126	DCEI	11
Online	71	Nasa	96	Nato	3
UMTS	32	Shuttle	35	Cluster	3
Handy	24	Crew	32	Manager	2
Ebay	24	Esa	23	Business	2

Table 3.2: Five most frequent (f) English inclusions per domain.

Table 3.2 also contains examples of English words with established and frequently used German equivalents such as *Crew* (Besatzung). The German translation of this term occurred 27 times in the space data. Therefore, the German word was used 45.8% and the English equivalent 54.2% of the time. English abbreviations such as *ISS* (International Space Station) or acronyms like *Esa* (European Space Agency) are specific cases of assimilated anglicisms as they are phonologically integrated in German.

A further interesting example listed in Table 3.2 is *Handy*, the word used by Germans for *mobile phone*. This is a pseudo-anglicism, 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 pseudo-anglicisms being 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*, for example, originated from the *Handy Talkie*, the first hand-held two-way radio developed in 1940 (Petrakis, 1965).

3.2.3 Inter-annotator Agreement

In any annotation project, some data is generally annotated by more than one annotator in order to guarantee consistency. Double (or multiple) annotation is also vital to determine how well defined a specific annotation task is, and how feasible it is for humans to perform. Inter-annotator agreement (IAA), which is calculated on a set of data annotated independently by different people, serves therefore as an upper bound of what is achievable by any system.

		Annotator A		
	Labels	English	Not English	Total
Annotator B	English	2,769	164	2,933
	Not English	381	93,385	93,766
	Total	3,150	93,549	96,699

Table 3.3: Contingency table for the English inclusion annotation.

In order to determine IAA figures, the entire German data set (development and test data) was annotated by 2 judges in parallel (annotator A and annotator B). The annotation guidelines are presented in Appendix B. IAA scores are calculated by means of a contingency table of the data versions produced by the annotators. The English inclusion annotation involves a binary annotation (English or Not English). The corresponding contingency table of both annotators for this task is shown in Table 3.3. For example, both annotators agreed on 2,769 tokens as being English and on 93,766 tokens as not being English. However, in 164 and 381 cases, one annotator marked the token as English whereas the other did not. Based on these figures, IAA scores can then be computed in terms of pairwise accuracy and F-score as well as the kappa coefficient, which are defined in Appendix A.2. The pairwise F-score for the English inclusion annotation of the two annotators is 91.04 and the accuracy amounts to 0.9944%. The κ-score is 0.9075 which equates to almost perfect agreement according to the criteria laid down by Landis and Koch (1977).

Since the IAA scores for annotating English inclusions are so high, it can be concluded that this task is not difficult for humans to perform. Analysing the annotation versions showed that some disagreement occurs for anglicisms like *Team*, *Job* or *Surfer*. These nouns have not entered German recently but are well established and widely used expressions. For other annotation projects, disagreements between two annotators are often resolved in an effort to create a reconciled corpus used for either training or evaluation purposes (e.g. Hachey et al., 2005). Due to time constraints and the relatively high agreement, this reconciliation phrase was dispensed with. Therefore, all evaluation figures reported in the remainder of this chapter are determined by comparing the system output to the annotations of one annotator only, annotator A.

3.2.4 Annotation Issues

Although the aforementioned annotation guidelines are relatively clear, the actual annotation revealed some tricky cases which were difficult to classify with the binary classification scheme described above. This section discusses the main issues which need to be clarified for revising and possibly extending the current guidelines.

These complicated instances mainly concern NEs which cannot be found in the individual lexicons but have certain language specific morphology and comprise character sequences typical for that language. Table 3.4 lists some examples for different types of NEs that stem from German, English and other language origins. *Dudenhöffer* and *Neckermann* are clearly German names, just as *Hutchison* and *Forrester* are English names. Difficult cases are *Sony* (sonus + sonny), *Activy* (similar to activity) or *Booxtra* (book + extra). Such English-like examples were not annotated as English inclusions in the gold standard. Therefore, if the system identifies them as English, its performance scores determined in the evaluation (Section 3.4) are to some extent unfairly penalised. A way of including these instances in the evaluation is to annotate them as English with an attribute distinguishing them from real English words.

NE type	German	English	Other
Person	Dudenhöffer	Hutchison	Kinoshita
Company	Neckermann	Forrester	Kelkoo

Table 3.4: Difficult annotation examples.

The German development corpus also contains NEs from other languages. Readers might well know that Kinoshita is neither a German nor an English name, although they may not be able to identify it as a Japanese name. Interestingly, Font-Llitjós and Black (2001) show that out of a list of 516 names, only 43% can be labelled confidently by human annotators with respect to their language origin. An example in our corpus where the language origin is not at all apparent in the character sequence is the name *Kelkoo*, a play on words derived from the French for *What a bargain* (Quel coup). It represents the English phonetic spelling of the French phrase. Other foreign names are *Toshiba* (Japanese), *Svanberg* (Swedish), *Altavista* (Spanish) and *Centrino* (Italian).

As yet, such instances are also annotated as O, i.e. not English. Entities from other languages could instead be annotated as INTERNATIONAL.

Another issue that arises is the annotation of place names. According to the annotation guidelines, English place names are not annotated as English inclusions unless they have a German equivalent which is generally used instead. So, for example, the location *Munich* would be annotated because the German equivalent *München* is generally the name preferred by German speakers. Examples of English location names found in the corpus that were not annotated are *Virginia*, *Houston* or *New York*. It can be argued that similar to the examples above they have certain English characteristics and sometimes coincide with common English words in the lexicon, as is the case for *Bath*. Identifying such examples, will be particularly advantageous for certain TTS applications such as car navigation systems where the correct pronunciation of place names is vital.

3.3 English Inclusion Classifier: System Overview

This section presents the **annotation-free English inclusion classifier** developed as part of this thesis project. It identifies English inclusions in German text by means of computationally inexpensive lexicon lookup and web search procedures. Annotation-free means that the system does not require any annotated training data and only relies on lexicons and the World Wide Web. This system allows linguists and lexicographers to observe language changes over time, and to investigate the use and frequency of foreign words in a given language and domain. The output also represents valuable information for a number of applications, including parsing, polyglot TTS synthesis and MT.

The initial system was built for a preliminary study 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. The system described in this section represents the final version which was developed on the basis of error analysis and numerous parameter tuning experiments. These fine-tuning experiments are described in detail in Section 3.5.

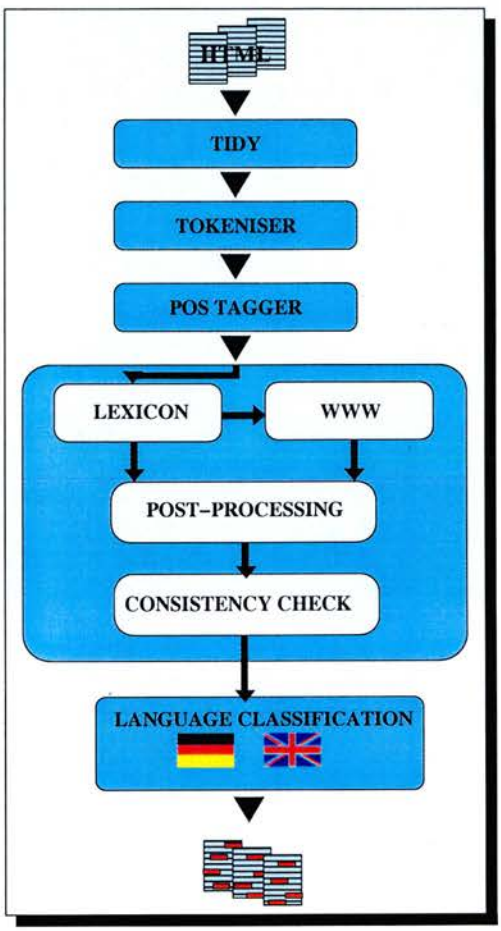


Figure 3.1: System architecture of the English inclusion classifier.

3.3.1 Processing Paradigm

The underlying processing paradigm of the English inclusion classifier is XML-based. As a markup language for NLP tasks, XML is expressive and flexible yet constrainable. Furthermore, there exists 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. As illustrated in Figure 3.1, the system for processing German text is essentially a UNIX pipeline which converts HTML files to XML and applies a sequence of modules: a preprocessing module for tokenisation and POS tagging, followed by a lexicon lookup, a search engine module, post-processing and an optional document consistency check which all add linguistic markup and classify tokens as either German or English. The pipeline is composed partly of calls to LT-TTT2 and LT-XML2 (Grover *et al.*, 2006)[4] for tokenisation and sentence splitting. In addition, non-XML public-domain tools such as the TnT tagger (Brants, 2000b) were integrated and their output incorporated into the XML markup. The primary advantage of this architecture is the ability to integrate the output of already existing tools with that of new modules specifically tailored to the task in an organised fashion. The XML output can be searched to find specific instances or to acquire counts of occurrences using the LT-XML2 tools.

3.3.2 Pre-processing Module

All downloaded Web documents are first of all cleaned up using TIDY[5] to remove HTML markup and any non-textual information and then converted into XML. Alternatively, the input into the classifier can be in simple text format which is subsequently converted into XML format. The resulting XML pages simply contain the textual information of each article. Subsequently, all documents are passed through a series of pre-processing steps implemented using the LT-XML2 and LT-TTT2 tools (Grover *et al.*, 2006) with the output of each step encoded in XML.

Two rule-based grammars which were developed specifically for German are used

[4]These tools are improved upgrades of the LT-TTT and LT-XML toolsets (Grover *et al.*, 2000; Thompson *et al.*, 1997) and are available under GPL as LT-TTT2 and LT-XML2 at: `http://www.ltg.ed.ac.uk`.

[5]`http://tidy.sourceforge.net`

to tokenise the XML documents. The first grammar pre-tokenises the text into tokens surrounded by white space and punctuation and the second grammar groups together various abbreviations, numerals and URLs. Grammar rules also split hyphenated tokens. The two grammars are applied with `lxtransduce`[6], a transducer which adds or rewrites XML markup to an input stream based on the rules provided. `lxtransduce` is an improved version of `fsgmatch`, the core program of LT-TTT (Grover *et al.*, 2000). The tokenised text is then POS-tagged using the statistical POS tagger TnT (Trigrams'n'Tags). The tagger is trained on the TIGER Treebank (Release 1) which consists of 700,000 tokens of German newspaper text (Brants *et al.*, 2002) annotated with the Stuttgart-Tübingen Tagset (Schiller *et al.*, 1995), henceforth referred to as STTS.

3.3.3 Lexicon Lookup Module

The lexicon module performs an initial language classification run based on a case-insensitive lookup procedure using two lexicons, one for the base language of the text and one for the language of the inclusions. The system is designed to search CELEX Version 2 (Celex, 1993), a lexical database of German, English and Dutch. The German database holds 51,728 lemmas and their 365,530 word forms and the English database contains 52,446 lemmas representing 160,594 corresponding word forms. A CELEX lookup is only performed for tokens which TnT tags as `NN` (common noun), `NE` (named entity), `ADJA` or `ADJD` (attributive and adverbial or predicatively used adjectives) as well as `FM` (foreign material). Anglicisms representing other parts of speech are relatively infrequently used in German (Yeandle, 2001) which is the principal reason for focussing on the classification of noun and adjective phrases. Before the lexicon lookup is performed, distinctive characteristics of German orthography are exploited for classification. So, all tokens containing German umlauts are automatically recognised as German and are therefore not further processed by the system.

The core lexicon lookup algorithm involves each token being looked up twice, both in the German and English CELEX databases. Each part of a hyphenated compound is checked individually. Moreover, the lookup in the English database is made case-insensitive in order to identify the capitalised English tokens in the corpus, the reason

[6] `http://www.ltg.ed.ac.uk/~richard/lxtransduce.html`

being 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 lexicon lookup, each token is found either: (1) only in the German lexicon, (2) only in the English lexicon, (3) in both or (4) in neither lexicon.

1. The majority of tokens found exclusively in the German lexicon are actual German words. Only very few are English words with German case inflection such as *Computern*. The word *Computer* is used so frequently in German that it already appears in lexicons and dictionaries. To detect the base language of inflected forms, a second lookup could be performed checking whether the lemma of the token also occurs in the English lexicon.

2. Tokens found exclusively in the English lexicon such as *Software* or *News* are generally English words and do not overlap with German lexicon entries. These tokens are clear instances of English inclusions and consequently tagged as such.

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

Table 3.5: Most frequent words per domain found in both lexicons.

3. Tokens which are found in both lexicons are words with the same orthographic characteristics in both languages (see Table 3.5). These are words without inflectional endings or words ending in *s* signalling either the German genitive singular case or the German and English plural forms of that token, e.g. *Computers*. The majority of these lexical items have the same or similar semantics in both languages and represent assimilated borrowings and cognates 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, namely Germanic languages, and have been influenced similarly by other foreign languages including Latin and French (Waterman, 1991). Only a small subgroup are clearly English borrowings (e.g. *Monster*). On the basis of careful error analysis, I designed a series of post-processing rules to disambiguate such English inclusions (Section 3.3.5). Some tokens found in both lexicons are interlingual homographs with different semantics in the two languages, e.g. *Rat* (*council* vs. *rat*). Deeper semantic analysis is required to distinguish the language of such homographs which are tagged as German by default at this point in the system. Moreover, it should be mentioned that English text contains some German loan words, though to a much lesser extent than vice versa. The German corpus contains such a relatively rare example, the word *Ersatz*, which is actually contained in the English lexicon.

4. All tokens found in neither lexicon include, for example:

   - German compounds, including loan substitutions: *Mausklick* (mouse click)
   - English unhyphenated compounds: *Homepage*, *Hypertext*
   - Mixed-lingual unhyphenated compounds: *Shuttleflug* (shuttle flight)
   - English nouns with German inflections: *Receivern* (with German dative plural case ending)
   - Abbreviations and acronyms: *UMTS*, *UKW*
   - Named entities: *Coolpix*, *Expedia*
   - English words with American spelling: *Center*
   - Words with spelling mistakes: *Abruch* (abort, correct spelling is *Abbruch*)
   - Other new German or English words that have not yet been entered into the dictionary: *Euro*, *Browser*

   Such ambiguous tokens which are not clearly identified by the lexicon module as either German or English are further processed by the search engine module described in the next section.

The results of evaluating the lexicon module as a separate component, as opposed to the overall system performance, are presented in Section 3.4.2.1.

3.3.4 Search Engine Module

The search engine module exploits the World Wide Web, a continuously expanding resource with textual material in a multiplicity of languages. Originally, the World Wide Web was a completely English medium. A study carried out by the Babel project[7] showed that in 1997 82.3% of a set of 3,239 randomly 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 rapidly (Crystal, 2001; Grefenstette and Nioche, 2000; Kilgarriff and Grefenstette, 2003). This increasing Web presence of languages can therefore be exploited as a rich and dynamic linguistic knowledge source.

The exploitation of the Web as a linguistic resource has become a growing trend in computational linguistics. Although the information published on the Web is sometimes noisy, its sheer size and the perpetual addition of new material make it a valuable pool of information in terms of languages in use. The Web has already been successfully exploited for several NLP tasks such as NE acquisition (Jacquemin and Bush, 2000), disambiguation of prepositional phrase attachments (Volk, 2001), anaphora resolution (Modjeska *et al.*, 2003), word sense disambiguation (Mihalcea and Moldovan, 1999; Agirre and Martinez, 2000) and MT (Grefenstette, 1999; Resnik, 1999). For a detailed overview of these experiments see also Keller and Lapata (2003).

The initial search engine module (Alex and Grover, 2004; Alex, 2005) was interfaced with the search engine Google. The principle motivation for this choice was the extremely large size of its search space: At the time, early 2004, Google had indexed more than 8 billion webpages, a large portion of all information available on the Web. Following a series of parameter tuning experiments described and discussed in Section 3.5, the search engine Yahoo is used now instead as it allows a larger number of automatic queries per day. Queries are submitted automatically by the search engine module via the Yahoo API. The module obtains 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 most search engines. So long as the search engine's internal language identification performs well, the underlying assumptions here is that a German word is more frequently used in German text than in English and vice versa.

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

The module therefore relies on the number of hits returned by the search engine as an indication of the actual frequency of the query in the documents accessible by the search engine. Each token is classified as either German or English based on the search that returns the maximum normalised score of the number of hits $rf_{C_{web}(L)}(t)$ returned for each language L. As shown in the following equation, this score is determined by weighting the number of hits, i.e. the "absolute frequency" $f_{C_{web}(L)}(t)$, by the size of the accessible Web corpus for that language, $N_{C_{web}(L)}$. The notation t designates token and C refers to corpus.

$$rf_{C_{web}(L)}(t) = \frac{f_{C_{web}(L)}(t)}{N_{C_{web}(L)}} \tag{3.1}$$

The size of the Web corpus for each language $N_{C_{web}(L)}$ is estimated following a method motivated by Grefenstette and Nioche (2000). $rf_{C_{std}}(w_{1...n})$, the relative frequencies of a series of common words within a standard corpus in a language, are used to make a series of n predictions on the overall size of the corpus of that language indexed by the search engine. This is done by dividing the actual number of hits of each word returned by the search engine by the relative frequency of the same word in the standard corpus. The total number of words in the particular language accessible through the search engine is then determined by taking the average of each individual word prediction:

$$N_{C_{web}(L)} = \frac{1}{n} \sum_{k=1}^{n} \frac{f_{C_{web}(L)}(w_k)}{rf_{C_{std}(L)}(w_k)} \tag{3.2}$$

Grefenstette and Nioche (2000)'s experiments were conducted with Altavista which at the time returned both page counts, the number of pages on which each query appears, as well as phrase counts, the number of times each query is indexed by Altavista. They regard the latter as an estimate of the actual frequency for each query in documents accessible by Altavista. As the phrase count feature has been discontinued both by Altavista and Yahoo and as Google only offer a total pages count, the only option is to rely on the latter figure for the present study. Zhu and Rosenfeld (2001) show that n-gram (unigram, bigram and trigram) page counts and phrase counts obtained from Altavista are largely log-linear and therefore highly correlated. This finding jus-

tifies the decision to use page counts instead of phrase counts as an estimate of actual Web frequencies. Moreover, the search engine module assumes that there is a close relationship between page counts and real corpus counts. Consequently, it is also vital to establish their correlation. Keller and Lapata (2003) demonstrate that bigram Web counts from Altavista and Google are highly correlated to corpus counts from the British National Corpus (BNC) and the North American News Text Corpus (NANTC). Their results also show that there is virtually no difference between the correlations determined using either search engine. This means that Web counts represent useful frequency information.

Two examples of how the search engine module identifies the language of a given word are presented in Figure 3.2 and Table 3.6. These searches were carried out in April 2006. At that time, the German Web corpus was estimated to contain approximately 53.3bn tokens and the English one around 638.9bn tokens, nearly 12 times as many as in the German Web corpus. The German word *Anbieter* (provider) occurred with an actual frequency of 62.0m and 11.2m in the German and English webpages indexed by Yahoo, respectively. Therefore, its weighted frequency in German Web documents (0.0116463) is considerably higher than that in the English Web documents (0.00001753). Conversely, the English equivalent, the word *provider*, occurs more often in English Web documents (168.0m) than in German Web documents (0.3m) resulting in a much higher weighted frequency in the English Web corpus (0.00026289) than the German one (0.00000626).

Figure 3.2: Yahoo queries with different language preferences.

Language preference	German		English	
Counts	Actual (f)	Normalised (rf)	Actual (f)	Normalised (rf)
Anbieter	62.0 M	**0.00116463**	0.333 M	0.00000626
Provider	11.2 M	0.00001753	168.0 M	**0.00026289**

Table 3.6: Actual and normalised frequencies of the search engine module for one German and one English example.

In the unlikely event that both searches return zero hits, the token is classified as the base language, in this case German, by default. In the initial experiment, this happened only for two tokens: *Orientierungsmotoren* (navigation engines) and *Reserveammoniak* (spare ammonia). Word queries that return zero or a low number of hits can also be indicative of new expressions that have entered a language.

The search engine module lookup is carried out only for the sub-group of tokens not found in either lexicon in the preceding module in order to keep the computational cost to a minimum. This decision is also supported by the evaluation of the lexicon module (Section 3.4.2.1) which shows that it performs sufficiently accurately on tokens contained exclusively in the German or English lexicons. Besides, current search options granted by search engines 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. *Rat*). The evaluation results specific to the search engine module as a separate component are presented in Section 3.4.2.1.

3.3.5 Post-processing Module

The final system component is a post-processing module that resolves several language classification ambiguities and classifies some single-character tokens. The post-processing rules are derived following extensive error analysis on the core English inclusion classifier output of the German development data. In the remainder of the thesis, the English inclusion classifier without post-processing is referred to as **core system** and with post-processing (and optional document consistency checking) as **full system**. The different types of post-processing rules implemented in this mod-

Post-processing type	Example
Ambiguous words	*Space **Station** Crew*
Single letters	***E**-mail*
Currencies & Units	***Euro***
Function words	*Friends **of the** Earth*
Abbreviations	*Europäische Union (**EU**)*
Person names	*Präsident **Bush***

Table 3.7: Different types of post-processing rules.

ule involve resolving language classification of ambiguous words, single letter tokens, currencies and units of measurement, function words, abbreviations and person names. Each type of post-processing is listed in Table 3.7 with an example and explained in more detail in the following. Individual contributions of each type are presented in Section 3.4.2.3. Most of the rules lead to improvements in performance for all of the three domains and none of them deteriorate the scores.

As only the token and its POS tag but not its surrounding context are considered in the lexicon module classification, it is difficult to identify the language of interlingual homographs, tokens with the same spelling in both languages (e.g. *Station*). Therefore, the majority of post-processing rules are designed to disambiguate such instances. For example, if a language ambiguous token is preceded and followed by an English token, then its is also likely to be of English origin (e.g. *Space Station Crew* versus *macht Station auf Sizilien*). The post-processing module applies rules that disambiguate such interlingual homographs based on their POS tag and contextual information.

Moreover, the module contains rules designed to flag single-character tokens correctly. These occur because the tokeniser is set up to split hyphenated compounds like *E-mail* into three separate tokens (Section 3.3.2). The core system identifies the language of tokens with a length of more than one character and therefore only recognises *mail* as English in this example. The post-processing rule flags *E* as English as well. Several additional rules deal with names of currencies and units of measurements and prevent them from being mistaken as English inclusions. Furthermore, some rules were designed to classify English function words as English.

As the core system classifies each token individually, a further post-processing

step is required to relate language information between abbreviations or acronyms and their definitions. These are firstly identified by means of an abbreviation extraction algorithms which functions based on character matching between short and long forms (Schwartz and Hearst, 2003). Subsequently, post-processing rules are applied to guarantee that each pair as well as earlier and later mentions of either the definition or the abbreviation or acronym are assigned the same language tag within each document.

Extensive error analysis also revealed that foreign person names (e.g. Hutchison) are frequently identified as English inclusions. This is not surprising as such tokens are likely not to be contained in the lexicons and when processed by the search engine module tend to have a higher relative frequency in English Web documents. At this point, the English inclusion classifier is merely evaluated on identifying actual inclusions. These are defined as English words and abbreviation except for person and location names (Section 3.2). Person names of English origin are not annotated in the English inclusion gold standard. To improve the performance of recognising real English inclusions, further post-processing rules are implemented to distinguish between the latter and English person names. The aim is to increase precision without reducing recall. Patterns signalling person names (e.g. 'Präsident X") were generated to distinguish them from English inclusions. Once a person name is identified all other mentions of it in the same document are also excluded. This system is therefore geared towards lexicographers who are more interested in the influx of English common words than in the mentioning of people's names. However, for a potential application of this system as a front-end to a TTS synthesis system, the additional language information of person names could prove beneficial for generating correct pronunciations.

After applying the post-processing rules described above, the balanced F-score on the German development data amounts to 82.17 points for the internet domain. The evaluation metric is defined in Appendix A.1. This represents an overall performance improvement of 5.59 points in F-score, 2.88% in precision and 5.99% in recall, over the core English inclusion classifier. The precision of the core system is already relatively high at 90.6%. The results for this and the other domains are examined in more detail in Section 3.4. The higher increase in recall shows that the post-processing is mainly aimed at identifying false negatives, i.e. ambiguous English inclusions missed by the core system. This supports the hypothesis that the language information of a token's surrounding context is highly beneficial to resolve ambiguities.

3.3.6 Document Consistency Checking

The English inclusion classifier is also designed to be combined with an optional consistency checking run in order to guarantee consistent classification within a given document. The consistency checking is designed to correct classification errors on the basis of wrong POS tag assignment. For example, the abbreviation *ISS* (International Space Station) is correctly classified as English when its POS tag is `NE`. However, whenever the POS tagger mistakes this token as a `VVPP` (perfect participle), the classifier is unable to identify it as English.

This particular problem was overcome by implementing a second classification run over the data using a gazetteer that is constructed on the fly during the first run. This means that whenever a token is classified as English by the full system, it is then added to the English inclusion gazetteer. After the first run, all tokens found in the English inclusion lexicon are tagged as English. This consistency checking is performed only on those tokens not already classified by the system in the first run. This allows for the classification of tokens which the system did not consider at first but at the same time avoids correcting decisions made earlier, for example in the post-processing module. This consistency checking is carried out on the document level. The motivation behind this decision is that repetitions of a specific interlingual homograph are likely to have the same meaning within a document but could have different semantics across documents. The evaluation of document consistency checking is presented and discussed in Section 3.4.2.4.

3.3.7 Output

The following are two example sentences of the system output retaining the English (EN) language classification alone for clarity. All the English inclusions in the first sentence, the headline of a newspaper article (Weiss, 2005), are correctly identified by the core English inclusion classifier. The lexicon module correctly identifies the compound noun *Security-Tool* as English. The noun phrases *Hacker*, *Google* and *Sicherheitslücken* are not listed in the lexicons and are therefore sent to the search engine module. It then correctly classifies the first two as English. The system also identifies *Sicherheitslücken* as German but since we already know that the base language of the sentence is German, this information is of less significance. The main goal is to iden-

tify the English inclusions in the utterance. In this case, all tokens are unambiguous and therefore no further post-processing is required.

```
<EN>Security</EN>-<EN>Tool</EN> verhindert, dass
<EN>Hacker</EN> über <EN>Google</EN>
Sicherheitslücken finden.
```

Translation: *Security Tool prevents hackers from finding security holes via Google.*

The second example is part of a quote made by fashion designer Jil Sander in an interview with FAZ (FAZ-Magazin, 1996). These words have become the prime example of anglicisation of German for which she was the first to receive the title "Language Diluter of the Year" from the Verein Deutscher Sprache e.V. (German language association) in 1997. This example contains numerous English inclusions, most of which are identified by the lexicon module (*contemporary*, *Future*, *Concept*, *Collection* and *Audience*). The tokens *Tailored*, *coordinated* and *supported* are correctly classified as English by the search engine module. The only ambiguous tokens are *Future* and *Hand*. They are resolved in the post-processing module on the basis of context.

```
Ich habe verstanden, daß man <EN>contemporary</EN> sein
muß, daß man <EN>Future</EN>-Denken haben muß.  Meine
Idee war, die <EN>Hand</EN>-<EN>Tailored</EN>-Geschichte
mit neuen Technologien zu verbinden.  Und für den Erfolg
war mein <EN>coordinated</EN> <EN>Concept</EN>
entscheidend, die Idee, daß man viele Teile einer
<EN>Collection</EN> miteinander combinen kann.  Aber
die <EN>Audience</EN> hat das alles von Anfang an auch
<EN>supported</EN>.
```

Translation: *I understood that one has to be contemporary, that one has to have future thinking. My idea was to combine the hand-tailored story with new technologies. And crucial to the success was my coordinated concept that one can combine parts of a collection. But the audience has supported this from the beginning as well.*

3.4 Evaluation and Analysis

This section first evaluates the performance of the English inclusion classifier on the German development data for each domain and subsequently examines the performance of individual system modules. Finally, the performance of the classifier on a random selection of unseen test data and another new data set provided by Marcadet *et al.* (2005) is reported. The latter allows comparison with another state-of-the-art mixed-lingual LID approach.

3.4.1 Evaluation of the Tool Output

The identification of English inclusions is similar to named entity recognition (NER) but on single tokens. The classifier's performance is therefore evaluated against the gold standard in terms of accuracy for all tokens, and balanced F-score (the harmonic mean of precision and recall) for target and predicted English tokens. Both metrics are defined in Appendix A.1. Baseline accuracy scores shown in Table 3.8 are determined assuming that the system found none of the English tokens in the data and believes that all tokens are German. As precision, recall and F-score are calculated in relation to the English tokens in the gold standard, they are essentially zero for the baseline. For this reason, only accuracy baseline scores (and not F-score baseline scores) are reported. Unsurprisingly, the baseline accuracies are relatively high as most tokens in a German text are German and the amount of foreign material is relatively small.

The full system, the combined lexicon lookup and search engine modules as well as post-processing and document consistency checking, yields relatively high F-scores of 84.37 and 91.35 for the internet and space travel data but only a low F-score of 66.67 for the EU data. The latter is due to the sparseness of English inclusions in that domain (see Table 3.1 in Section 3.2). Although the recall for this data (76.19%) is comparable to that of the other two domains, the number of false positives is high, causing low precision and F-score. Results were compared using the chi square (χ^2) test (see Appendix A.4.1). It shows that the additional classification of English inclusions yields highly statistically significant improvements ($df = 1$, $p \leq 0.001$) in accuracy over the baseline of 4.30% for the internet data and 2.46% for the space travel data. When classifying English inclusions in the EU data, accuracy increases only slightly by 0.09% which is not statistically significant ($df = 1$, $p \leq 1$).

Domain	Method	Accuracy	Precision	Recall	F-score
Internet	Baseline	93.95%	-	-	-
	Full system	98.25%	92.75%	77.37%	84.37
	TextCat	92.24%	33.57%	28.87%	31.04
Space	Baseline	96.99%	-	-	-
	Full system	99.45%	89.19%	93.61%	91.35
	TextCat	93.80%	20.73%	37.32%	26.66
EU	Baseline	99.69%	-	-	-
	Full system	99.78%	59.26%	76.19%	66.67
	TextCat	96.43%	2.54%	28.57%	4.66

Table 3.8: Performance of the English inclusion classifier compared to the baseline and the performance of TextCat.

In order to get an idea of how a conventional LID system performs on the task of recognising English inclusions embedded in German text, Table 3.8 also reports the performance of TextCat, an automatic LID tool based on the character n-gram frequency text categorisation algorithm proposed by Cavnar and Trenkle (1994) and reviewed in Section 2.2. While this LID tool requires no lexicons, its F-scores are low for the internet and space travel domains (31.04 and 26.66, respectively) and very poor for the EU data (4.66). This confirms that the identification of English inclusions is more difficult for this domain, coinciding with the result of the English inclusion classifier. The low scores also prove that such conventional n-gram-based language identification alone is unsuitable for token-based language classification, particularly in case of closely related languages.

3.4.2 Evaluation of Individual System Modules

The full system described in Section 3.3 combines a lexicon lookup module, a search engine module and a post-processing module in order to classify English inclusions in German text. This section reports the performance of individual system modules of the English inclusion classifier compared to those of the full system and the baseline scores. It shows that the combination of individual models leads to a performance increase of the system on mixed-lingual data.

3.4.2.1 Evaluation of the Lexicon and Search Engine Modules

In the first experiment, the system is limited to the lexicon module described in detail in Section 3.3.3. Lexicon lookup is restricted to tokens with the POS tags NN, NE, FM, ADJA and ADJD. Post-processing and document consistency checking, as carried out in the full system and described in Sections 3.3.5 and 3.3.6, are not applied here. Therefore, ambiguous tokens found in neither or both databases are considered not to be of English origin by default. The assumption is that the lexicon module performs relatively well on known words contained in the lexicons but will disregard all tokens not found in the lexicons as potential English inclusions. Therefore, precision is expected to be higher than recall. In the second experiment, the system is restricted to the search engine module only. Here, all tokens (with the POS tags NN, NE, FM, ADJA and ADJD) are classified by the search engine module based on the number of normalised hits returned for each language. Exact details on how this module functions are presented in Section 3.3.4. This experiment also does not involve any post-processing. As all queried tokens are treated as potential English inclusions, recall is expected to increase. Since some tokens are named entities which are difficult to classify as being of a particular language origin, precision is likely to decrease.

As anticipated, recall scores are low for the lexicon-only-evaluation across all domains (Internet: R=23.04%, Space: R=28.87%, EU: R=38.10%). These are due to the considerable number of false negatives, i.e. English inclusions that do not occur in the lexicon (unknown words). Conversely, Table 3.9 shows higher precision values for the lexicon module across all three domains (Internet: P=90.57%, Space: P=77.78%, EU: P=47.06%). In the search engine module evaluation, recall scores improve considerably, as expected (Internet: R=81.02%, Space: R=97.11%, EU: R=88.10%). On the other hand, this latter setup results in much lower precision scores (Internet: P=68.82%, Space: P=40.71%, EU: P=6.99%) which is partly due to the fact that Yahoo, as most search engines, is not sensitive to linguistic and orthographic information such as POS tags or case. For example, the German noun *All* (space) is classified as English because the search engine mistakes it for the English word "all" which is much more commonly used on the internet than its German homograph. Interlingual homographs are therefore often wrongly classified as English when running the search engine module on its own.

Domain	Method	Accuracy	Precision	Recall	F-score
Internet	Baseline	93.95%	-	-	-
	Lexicon module	95.09%	90.57%	23.04%	36.74
	Search engine module	96.60%	68.82%	81.02%	74.43
	Core system	97.47%	90.60%	66.32%	76.58
	Full system	98.25%	92.75%	77.37%	84.37
Space	Baseline	96.99%	-	-	-
	Lexicon module	97.57%	77.78%	28.87%	42.11
	Search engine module	95.62%	40.71%	97.11%	57.37
	Core system	99.05%	84.85%	84.33%	84.59
	Full system	99.45%	89.19%	93.61%	91.35
EU	Baseline	99.69%	-	-	-
	Lexicon module	99.69%	47.06%	38.10%	42.11
	Search engine module	96.94%	6.99%	88.10%	12.96
	Core system	98.41%	10.57%	66.67%	18.24
	Full system	99.78%	59.26%	76.19%	66.67

Table 3.9: Evaluation of the lexicon and search engine modules compared to the core and full systems as well as the baseline.

The core English inclusion classifier essentially combines a high precision lexicon module with a high recall search engine module. This is achieved by first running the lexicon module to classify all known words. Subsequently, the search engine module only processes unknown words, namely those tokens that are not classified by the lexicon module. A token which is not resolved by this combined classification process is considered not to be an English inclusion by default. The combined core system outperforms the individual lexicon and search engine modules both for the internet and space travel data with F-scores of 76.58 and 84.59, respectively. Both of these domains contain considerably large numbers of English inclusions. For the EU data, which only contains very few English inclusions, the lexicon module was only outperformed by the full system due to the additional post-processing. One of the errors that seriously decreased the performance of the core system was made by the search engine module. It recognised the abbreviation *EU* as English. This error was corrected by means of abbreviation post-processing described in Section 3.3.5.

Compared to the core system, the full English inclusion classifier involves a final post-processing stage as well as a document consistency check which are evaluated in more detail in Sections 3.4.2.3 and 3.4.2.4. The full system resulted in overall best F-scores for all three domains (Internet: F=84.37, Space: F=91.35, EU: F=66.67).

3.4.2.2 Web Search versus Corpus Search

In order to understand the merit of the search engine module and the amount of data it can access better, the search engine module was replaced with a corpus search module that determines relative token frequencies based on fixed corpora. Here, the language classification is essentially based on real corpus frequencies rather than estimated web corpus frequencies. Language identification is simply conducted as a result of the higher relative frequency (rf) of a token (t) for a given corpus (C) in a particular language (L) and calculated as the actual frequency of a token in the corpus normalised by the corpus size (N).

$$rf_{C(L)}(t) = \frac{f_{C(L)}(t)}{N_{C(L)}} \tag{3.3}$$

If the relative frequency of the token in the English corpus is higher than that in the corpus of the base language of the text, the token is classed as English. This experimental setup therefore requires two corpora, one for the inclusion language (English) and one for the base language of the text (German). In the initial experiment, two corpora of roughly equal size were used: the Wall Street Journal section of the Penn Treebank corpus, Version 3.0 (Marcus *et al.*, 1993) amounting to around 1.2m tokens and the combined German NEGRA and TIGER corpora (Skut *et al.*, 1998; Brants *et al.*, 2002) containing approximately 1.1m tokens. Both data sets were published in the 1990s. For the purpose of determining the relative frequencies of a given token for both languages and identifying its language accordingly, the corpora were converted into frequency lists.

All subsequent corpus search experiments are conducted using the German development set of newspaper articles in the internet & telecoms domain, the set containing the highest percentage of English inclusions. The architecture of the classifier is essentially the same as that of the English inclusion classifier, except that the search engine module is replaced by the corpus search module. Relative token frequencies are calculated using the same equations as in the search engine module, but based on a fixed corpus, instead of an estimated Web corpus for each language. The corpus search engine module is preceded by the pre-processing and lexicon modules and followed by optional post-processing. Document consistency checking is not applied.

As can be seen in Table 3.10, using the Wall Street Journal corpus as the basis for language identification in the corpus search module only increases the performance of the English inclusion classifier by 9.36 points in F-score compared to running the lexicon module alone. This score is far from the performance achieved with the combined lexicon and search engine module (F=76.58). The relatively poor result of the corpus search module is partially caused by the fact that the English Wall Street Journal corpus is limited in size and may therefore not cover the English terms that occur in the articles belonging to the German development set. Conversely, the likelihood that a word is not found online is very small given that search engines have access to billions of words. The other reason for the low score is the time period during which the Wall Street Journal corpus was published (1993-1994). While this English corpus is a relatively old collection, the German internet newspaper articles were published more recently between 2001 and 2005. It is therefore extremely likely that the English

Corpus Size	No. of Types	F-score without PP	F-score with PP
Lexicon module only			
N/A		36.74	39.11
Lexicon + corpus search module: Wall Street Journal corpus			
1,173,747	43,808	46.10	48.64
Lexicon + search engine module			
638.9bn tokens (estimate)[8]		76.58	82.17

Table 3.10: Evaluation of the corpus search module using the Wall Street Journal corpus and the combined NEGRA/TIGER corpus with/without post-processing (PP) compared to the lexicon module only and a combined lexicon and search engine module approach.

inclusions, which to some extent are recently emerged technological and computing vocabulary, did not exist or were not commonly used in the early 1990s. Moreover, unlike the German development set, the Wall Street Journal corpus contains general newspaper text not limited to a specific topic. This discrepancy in domain is another crucial factor in the small performance increase of combining the corpus search module with the lexicon module.

The corpus search module is set up to test the hypothesis that the search engine module performs better due to the large amount of data it can access, and the fact that this data is constantly updated and increased with new material. The aim is to simulate the search engine module behaviour in a more controlled fashion by making use of increasing corpus sub-sets. These are drawn from a corpus more recently released than the Wall Street Journal corpus, the Agence France Presse content of the English Gigaword corpus[9] (published between 1994-1997 and 2001-2002). The corpus sub-sets are created by randomly selecting sentences from the Gigaword corpus amounting to 1m, 10m, 20m, 30m and 40m tokens. While the German corpus (combined NEGRA/TIGER) remains unchanged, each of the English corpus sub-sets are used by the corpus search module in a separate run of the classifier over the German development data. The idea is to grant the corpus search module access to more and more data to

[8]The English web corpus estimation was carried out in April 2006.

[9]http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05

Corpus Size	Avg No. of Types	F-score without PP	F-score with PP
Lexicon module only			
N/A		36.74	39.11
Lexicon + corpus search module: Gigaword corpus			
1,000,000	52,268	60.37	67.06
10,000,000	165,445	65.41	71.92
20,000,000	229,139	66.73	73.18
30,000,000	273,139	69.74	74.74
40,000,000	308,421	70.89	75.87
Lexicon + search engine module			
638.9bn tokens (estimate)		76.58	82.17

Table 3.11: Evaluation of the corpus search module with increasing sub-sets of the Gigaword corpus with/without post-processing compared to the lexicon module only and a combined lexicon and search engine module approach.

identify the language of individual tokens.

Table 3.11 reports the F-scores with and without post-processing averaged over 5 repeated runs using a different selection of Gigaword sentences each time. In order to simulate the availability of increasingly larger data sets to the corpus search module, the amount of tokens extracted from the English Gigaword corpus is increased incrementally from 1m up to 40m tokens. Results are listed with and without post-processing for increasing corpus sizes. As expected, granting the corpus search module access to larger amounts of data results in an incremental performance increase in F-score. Using an English corpus of 1m tokens, the corpus search module results in an F-score of 60.37, i.e 23.63 points higher than when just applying the lexicon module and 16.21 points lower than when using the search engine module in its place. Given that this F-score is 14.27 points higher compared to the one obtained when using the Wall Street Journal (almost equal in size) this shows that data currentness is vital for English inclusion detection. The classifier improves steadily with access to larger corpus frequency lists and reaches an F-score of 70.89 when the corpus search module determines relative token frequencies in an English corpus containing 40m tokens.

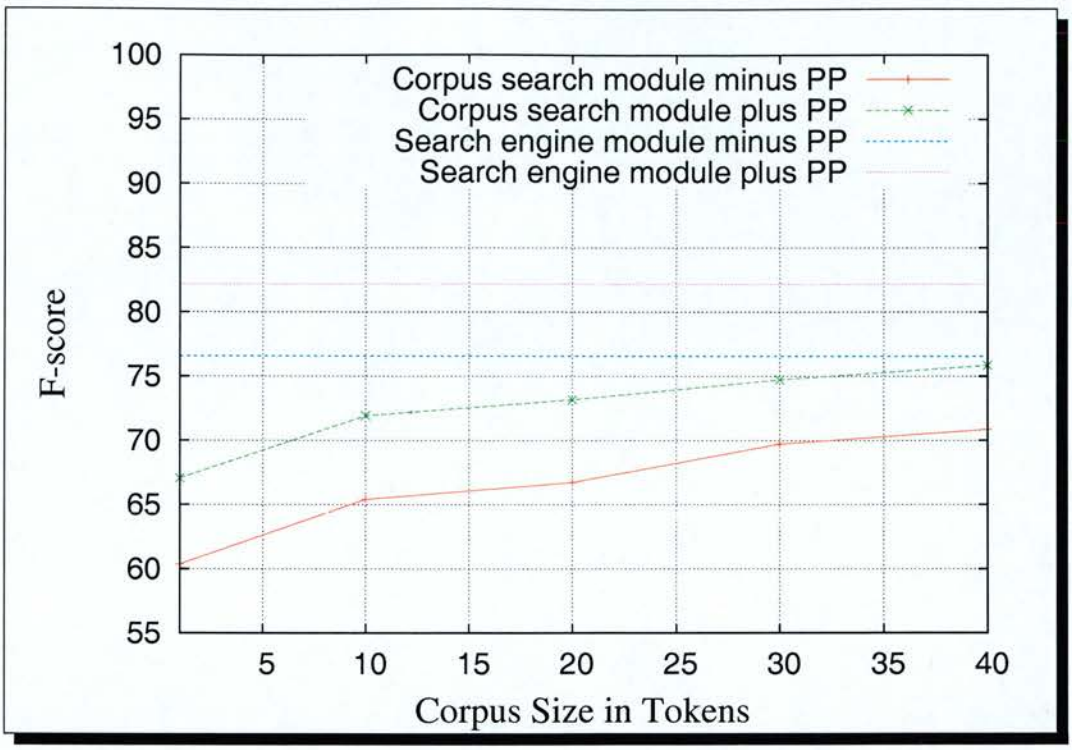


Figure 3.3: Performance increase of the corpus search module (with/without post-processing) with increasing sub-sets of the Gigaword corpus, compared to the search engine module's performance (with/without post-processing) represented as horizontal lines.

Figure 3.3 shows that the performance increases are reduced with larger corpus sizes.

To summarise, it was shown that token-based language identification improves with access to larger data sets. It also emerged that the time of publishing is an important aspect that needs to be considered. The use of any fixed-size corpus for language identification purposes clearly has its drawbacks. Such a collection is unlikely to contain all possible lexical items and, with languages evolving constantly, is out-of-date as soon as it is created and made available. Search engines provide access to extremely large collections of data which are constantly updated and changing with time and language use. Therefore, the search engine module has a clear superiority over accessing a corpus that is a data snap-shot of a particular time period and is limited in size. This is clearly reflected in the performance comparison of both methods. Access to a considerably larger corpus would be required for the corpus search module to reach the same level of performance as that of the search engine module.

3.4.2.3 Evaluation of the Post-Processing Module

The post-processing module yields a considerable improvement in F-score over the core English inclusion classifier for all three domains (see Table 3.9). This section provides an overview of the improvement gained from individual post-processing rules described in Section 3.3.5. Table 3.12 presents lesion studies showing the individual contribution of each post-processing rule to the overall performance of the full English inclusion classifier (without consistency checking) on the German development data. In this case, the term lesion study refers to eliminating a type of post-processing in order to examine its effect on the whole system. This type of experiment is also referred to as ablation study. The performance gain in F-score resulting from applying each type of post-processing is listed in the last column of Table 3.12. While some of the post-processing rules are specific to a particular data set with improvements of varying degree, none of them decreased the overall performance when applied.

The post-processing rules resulting in the largest performance increase are those designed to resolve the language of ambiguous words, single-character tokens and person names. The improvement of single-character token post-processing is particularly high for the internet domain as this data set contains frequent E-words like *E-Mail* or *E-Business*. While this rule leads to a small improvement for the space travel domain, it does not improve the performance for the EU domain as this data set does not contain such words. The rules disambiguating person names from real English inclusions yield large improvements for the EU and space travel data, as these data sets contain many foreign person names. The post-processing step which handles abbreviations, acronyms and their definitions leads to small improvements for the internet and space travel data but strongly increases the F-score for the EU data. This is due to the fact that the core system wrongly classified the token *EU* (short for *Europäische Union*) as English which occurs extremely frequently in this data set. Overall smaller improvements in F-score for all three domains result from the post-processing rules designed to disambiguate function words, currencies and units of measurements.

In total, post-processing results in a non-negligible performance increase for the internet and space travel data (3.27 and 4.77 points in F-score, respectively) and an extremely large improvement of 46.98 points in F-score for the EU data.

Post-processing	Accuracy	Precision	Recall	F-score	Δ F
Internet					
None	97.47%	90.60%	66.32%	76.58	3.27
Single letters	97.81%	93.23%	68.93%	79.26	2.91
Ambiguous words	97.91%	93.56%	71.22%	80.88	1.29
Person names	98.00%	92.02%	73.31%	81.60	0.57
Currencies etc.	98.02%	92.50%	73.31%	81.79	0.38
Abbreviations	98.04%	92.98%	73.20%	81.91	0.12
Function words	98.07%	93.48%	73.20%	82.11	0.07
Full System - CC	98.07%	93.48%	73.31%	82.17	-
Space					
None	99.05%	84.85%	84.33%	84.59	4.77
Person names	99.15%	85.14%	87.42%	86.27	3.09
Single letters	99.33%	91.30%	86.60%	88.89	0.47
Ambiguous words	99.34%	91.68%	86.39%	88.96	0.40
Abbreviations	99.33%	91.32%	86.80%	89.01	0.35
Function words	99.35%	91.36%	87.22%	89.24	0.12
Currencies etc.	99.35%	91.18%	87.42%	89.26	0.10
Full System - CC	99.36%	91.38%	87.42%	89.36	-
EU					
None	98.41%	10.57%	66.67%	18.24	46.98
Abbreviations	98.56%	12.24%	71.43%	20.91	44.31
Person names	99.64%	41.67%	71.43%	52.63	12.59
Ambiguous words	99.76%	58.33%	66.67%	62.22	3.00
Single letters	99.78%	60.00%	71.43%	65.22	0
Function words	99.78%	60.00%	71.43%	65.22	0
Currencies	99.78%	60.00%	71.43%	65.22	0
Full System - CC	99.78%	60.00%	71.43%	65.22	-

Table 3.12: Evaluation of the post-processing module with one rule removed at a time on the German development data. Δ F represents the change in F-score compared to the full English inclusion classifier without consistency checking (CC).

3.4.2.4 Evaluation of Document Consistency Checking

Table 3.13 shows the improvements in F-score obtained when adding document-based consistency checking (CC) to the English inclusion classifier. They amount to 2.2 points in F-score for the internet data, 1.99 points for the space travel data and 1.45 points for the EU data. This setup yields overall best F-scores for all three domains (Internet: F=84.27, Space: F=91.35, EU: F=66.67). It should be noted that this performance increase can be attributed to the rise in recall. After applying CC, all precision scores are marginally lower than those of the full classifier. While the overall improvement is essential for document classification, e.g. when comparing different classifiers as is done in the next section, it may not be beneficial for language classification on tokens in individual sentences, e.g. during the text analysis of a TTS synthesis system. In fact, the utility of document consistency checking is highly application dependent.

Method	Accuracy	Precision	Recall	F-score
Internet				
Baseline	93.95%	-	-	-
Full system - CC	98.07%	93.48%	73.31%	82.17
Full system + CC	98.25%	92.75%	77.37%	84.37
Space				
Baseline	96.99%	-	-	-
Full system - CC	99.36%	91.38%	87.42%	89.36
Full system + CC	99.45%	89.19%	93.61%	91.35
EU				
Baseline	99.69%	-	-	-
Full system - CC	99.78%	60.00%	71.43%	65.22
Full system + CC	99.78%	59.26%	76.19%	66.67

Table 3.13: Full system plus/minus consistency checking (CC) versus the baseline.

3.4.3 Evaluation on Unseen Data

All previously presented evaluation was carried out on the development set for each domain. The final system design of the English inclusion classifier is the result of various adjustments made after extensive error analysis and parameter tuning, described in detail in Section 3.5. It is therefore necessary to evaluate the system on entirely unseen data in order to determine its real performance. In the following, the results of an evaluation using such unseen data (Section 3.4.3.1) as well as a new data set provided by another research group (Section 3.4.3.2) are reported and discussed. The first data set was randomly selected which means that evaluation on that set represents the English inclusion classifier's performance on running newspaper text of various domains. The second data set was collected specifically for the task of mixed-lingual LID. Evaluation using this latter set will determine the performance of the system compared to another state-of-the-art approach taken by Marcadet *et al.* (2005).

3.4.3.1 Unseen Test Data

First, it is of particular interest to test how well the English inclusion classifier performs on completely unseen data in all three domains. For this purpose, a manually annotated test data set for each domain of approximately equal size as the development set was used (see Section 3.2 for details on data preparation). The results in Table 3.14 illustrate how well the full English inclusion classifier performs on this unseen test data for all three domains. For ease of comparison, the results for the development data are presented as well. The table lists the result of the full system with optional document consistency checking (see Section 3.3.6).

Overall, the full system F-scores for the test data are relatively high across all three domains, ranging between 82 and 85 points, which means that the classifier performs well on new data. This constitutes an advantage over supervised machine learning (ML) methods which require constant retraining on new annotated training data. The performance of a supervised maxent classifier on identifying English inclusions will be investigated further in Section 3.6. Interestingly, the F-score for the test data in the internet domain is approximately 1 point higher than that for the internet development data without document consistency checking (83.18 versus 82.17 points). This difference is reduced to 0.41 when consistency checking is applied.

	Test set				Development set			
Method	Acc	P	R	F	Acc	P	R	F
Internet								
BL	93.62%	-	-	-	93.95%	-	-	-
FS	97.93%	92.13%	75.82%	83.18	98.07%	93.48%	73.31%	82.17
FS+CC	98.13%	91.58%	78.92%	84.78	98.25%	92.75%	77.37%	84.37
Space								
BL	97.19%	-	-	-	96.99%	-	-	-
FS	98.89%	85.61%	79.61%	82.50	99.36%	91.38%	87.42%	89.36
FS+CC	98.97%	84.02%	85.31%	84.66	99.45%	89.19%	93.61%	91.35
EU								
BL	98.93%	-	-	-	99.69%	-	-	-
FS	99.65%	83.24%	85.63%	84.42	99.78%	60.00%	71.43%	65.22
FS+CC	99.65%	82.16%	87.36%	84.68	99.78%	59.26%	76.19%	66.67

Table 3.14: Evaluation of the full system (FS) on the unseen test data with optional consistency checking (CC) versus the baseline (BL).

The F-score for the space travel test data is almost 7 points lower than that obtained for the development set. This performance drop is caused by lower precision and recall. Although the classifier is overgenerating on the development set for this particular domain, the fact that the scores on the unseen test data are relatively consistent across all three different domains is a positive result. Moreover, each data set is relatively small which makes it difficult to draw clear conclusions. In fact, the test and development data on space travel are slightly different in nature as can be seen in Table 3.1. While both sets contain a similar percentage of English inclusions (2.8% versus 3%), those in the test set are much less repeated than those in the development set which is reflected in their type-token-ratios amounting to 0.33 and 0.15, respectively. Therefore, the higher development test data scores could be due to the higher number of repetitions of English inclusions in the space travel development data.

The full system F-scores for the EU test data are considerably higher than for the development set (84.42 versus 65.22 points). This is not surprising since the EU development data only contains 30 different English inclusions, less than 1% of all types, which made it an unusual data set for evaluating the classifier on. Error analysis was therefore focused mainly on the output of the other two data sets. The EU test data, on the other hand, contains 86 different English inclusions, i.e. three times as many types as in the development data (see Table 3.1). Considering that the English inclusion classifier yields an equally high performance on the unseen EU test data as on the other two test data sets, it can be concluded that system design decisions and post-processing rules are made general enough to apply to documents on different domains.

The best overall F-scores on all six data sets are obtained when combining the full system with a second consistency checking run (Internet test data: F=84.78, Space travel test data: F=84.66, EU test data: F=84.68). This second run essentially ensures that all English inclusions found in the first run are consistently classified within each document. This is done by applying an on-the-fly gazetteer which is generated automatically. This setup was explained in more detail in Section 3.3.6. The results listed in Table 3.14 show that the improvement in F-score is always caused by an increase in recall, outweighing the smaller decrease in precision. While this improvement is essential for document classification, particularly when comparing different classifiers, it is unlikely to be beneficial when performing language classification on tokens in individual sentences, for example during the text analysis of a TTS synthesis system.

Evaluating the English inclusion classifier on the entire German data, i.e. the pooled development and test sets for all three domains, yields an overall F-score of 85.43 and an accuracy of 99.42%, 2.68% better than the baseline of assuming that every token is not English. Table 3.15 also shows that the classifier's performance is approaching the inter-annotator agreement (IAA) score which represents the upper bound for English inclusion detection. Examining the scores obtained when applying TextCat (see Section 2.2), a conventional LID tool, to this task shows that it is not suitable for token-based LID. Its accuracy of 93.65% is considerably lower than the baseline.

Method	Accuracy	Precision	Recall	F-score
English inclusion classifier	99.42%	89.08%	82.06%	85.43
IAA	99.44%	94.41%	87.90%	91.04
TextCat	93.65%	20.84%	34.70%	26.04
Baseline	96.74%	-	-	-

Table 3.15: Evaluation of the English inclusion classifier on the entire German data (development + test data) compared to the IAA, TextCat and the baseline.

3.4.3.2 Marcadet *et al.* (2005) Data

Language identification for polyglot speech synthesis became the focus of the IBM research group who developed a language detector designed as a front-end to a polyglot TTS system (Marcadet *et al.*, 2005). Their system and its performance is described in detail in Section 2.2.1.2. It identifies the language of input sentences on the token level and is designed to differentiate between English, French, German, Italian and Spanish. Marcadet *et al.* (2005) evaluated their system on three test sets, including a German one which is hand-annotated for English inclusions. The German test script is made up of 1,050 tokens of which 123 are English. The English inclusions therefore make up 11.71% of all tokens (and 15.73% of all types) in this data set. As is reflected in these high percentages, this data set is not a random selection of newspaper text but consists of sentences which were specifically extracted for containing English inclusions.

According to the gold standard annotation, this data set does not contain words of French, Italian or Spanish origin. The ten most frequent English inclusions in this data set are listed in Table 3.16. They are either proper or common nouns, adjectives or

German Test Corpus			
Token	f	Token	f
Windows	10	of	2
Microsoft	4	National	2
Word	3	Motherboards	2
Controller	3	Center	2
Rockets	2	Celeron	2

Table 3.16: Ten most frequent (f) English inclusions in the German test corpus of Marcadet *et al.* (2005).

prepositions which the English inclusion classifier, presented in this thesis, is specifically designed to deal with. With the availability of this test corpus, the opportunity arose to evaluate the performance of the English inclusion classifier on an entirely new data set. This not only facilitates determining the performance of the English inclusion classifier on a data set designed specifically for its purpose and comparing it against that of the system developed by Marcadet *et al.* (2005), it also makes it possible to compute IAA for marking up English inclusions in German text.

In order to determine the IAA, all language markup was removed from the data which was then re-annotated according to the annotation guidelines presented in this thesis (Appendix B). Commonly used metrics for measuring inter-annotator agreement are pairwise F-score and the κ coefficient (see Appendix A.2). According to the latter metric, which represents agreement beyond what is expected by chance, the two annotators identify English inclusions in German text with a reliability of κ=.844, indicating almost perfect agreement. However, when evaluating my annotation against Marcadet et al.'s original annotation in terms of precision, recall and F-score in order to identify instances of annotation disagreement, the differences become more apparent. The F-score only amounts to 86.26 points (81.29% precision and 91.87% recall). Although computed on much less data, this IAA F-score is 4.78 points lower than that obtained for the doubly annotated German data created as part of this thesis project (see Section 3.2.3). It is also relatively low considering that most educated Germans are expected to identify English inclusions in German text without serious problems. For example von Wickert (2001) shows in a survey that English notions and slogans

occurring in advertising are conspicuous for nearly all of the respondents (98%). A closer look at the data reveals that Marcadet *et al.* (2005) do not consistently annotate abbreviations and acronyms expanding to English definitions as English. Conversely, the English inclusion classifier presented in this thesis is designed to recognise them as well. Moreover, person names like *Ted Saskins* are annotated as English and not distinguished from real English inclusions as advocated in this thesis.

On the reconciled gold standard, the English inclusion classifier performs with an F-score of 96.35 points (an accuracy of 98.95%, a precision of 97.78% and a recall of 94.96%). These scores are slightly better than those reported by Marcadet *et al.* (2005) on this data set (98.67% accuracy). However, it is not entirely straightforward to compare these scores as the gold standard annotation is reconciled. The few classification errors are mainly due to English words like *Team* or *Management* being already listed in the German lexicon. These anglicisms are strongly integrated in the German language and have well established pronunciations. Therefore, such classification errors are unlikely to cause pronunciation problems during TTS synthesis.

Given the results of both sets of evaluations, it can be concluded that the English inclusion classifier performs well on randomly selected unseen mixed-lingual data in different domains and compares well to an existing mixed-lingual LID approach.

3.5 Parameter Tuning Experiments

This section discusses a series of interesting parameter tuning experiments to optimise the English inclusion classifier. These were the basis for the final design of the full system which was evaluated in the previous section. These experiments include a task-based evaluation of three different POS taggers and a task-based evaluation of two search engines. All experiments involve the German development data for evaluation.

3.5.1 Task-based Evaluation of Different POS taggers

Throughout the entire process of error analysis, it was noticed that the performance of the English inclusion classifier depends to some extent on the performance on the POS tagger. Initially, the system made use of the POS tagger TnT (Brants, 2000b) trained on the NEGRA corpus (Skut *et al.*, 1997). Some classification errors result from errors

made by the POS tagger and therefore could be avoided if the latter performed with perfect accuracy. One reason for lower tagging accuracy is the fact that POS taggers trained on data for a particular language do not necessarily deal well with text containing foreign inclusions. Moreover, some taggers have difficulty differentiating between common and proper nouns in some cases.

In order to gain a better understanding of how the POS tagging influences the performance of the English inclusion classifier, I compared three different taggers in a task-based evaluation:

- TnT_{NEGRA} - the TnT tagger trained on the NEGRA corpus of approximately 355,000 tokens (Skut *et al.*, 1997)
- TnT_{TIGER} - the TnT tagger trained on the TIGER corpus of approximately 700,000 tokens (Brants *et al.*, 2002)
- TreeTagger - the TreeTagger trained on a small German newspaper corpus of Stuttgarter Zeitung containing 25,000 tokens (Schmid, 1994, 1995)

The English inclusion classifier is essentially run on the same set of data tagged by the three different POS taggers and evaluated against the hand-annotated gold standard. Note that this method does not necessarily determine the best and most accurate POS tagger but rather one that is most beneficial for identifying English inclusions in German text. Before discussing the results for each setup, the characteristics of the two POS taggers used in this evaluation are explained in detail.

3.5.1.1 TnT - Trigrams'n'Tags

TnT is a very efficient statistical POS tagger which can be trained on corpora in different languages and domains and new POS tag sets (Brants, 2000b). Moreover, the tagger is very fast to train and run. It is based on the Viterbi algorithm for second order Markov models and therefore assigns the tag t_i that is most likely to generate the w_i given the two previous tags t_{i-1} and t_{i-2}. The output and transition probabilities are estimated from an annotated corpus. In order to deal with data sparseness, the system incorporates linear interpolation-based smoothing and handles unknown words via n-gram-based suffix analysis.

In 10-fold cross-validation experiments carried out by Brants (2000b), TnT performs with an accuracy of 96.7% on the Penn Treebank (1.2 million tokens) which represents a near state-of-the-art performance for English text. The tagger only yields an accuracy of 94.5% on the English Susanne corpus (150,000 tokens) which is unsurprising given that this data set is much smaller in size and is annotated with a much larger POS tag set (over 160 tags). The tagger's performance on the German NEGRA corpus of 96.7% accuracy is high.

The difference between the two versions of the TnT tagger used in the following experiments (TnT_{NEGRA} and TnT_{TIGER}) is merely that the first is trained on half the amount of training material (the NEGRA corpus) compared to the second (the TIGER corpus) but using a similar set of POS tags (see Appendix C).[10] Note that both corpora are distinct sets and do not overlap in parts. The aim is to test whether a tagger trained on less sparse data will improve the performance of classifying English inclusions.

3.5.1.2 TreeTagger

The TreeTagger, on the other hand, is based on decision trees for annotating text with POS and lemma information. The off-the-shelf implementation is distributed with models trained on German, English, French, Italian, Greek and ancient French text and can be adapted to other languages given the availability of a lexicon and a manually tagged training corpus. The TreeTagger estimates transition probabilities of n-grams by means of a binary decision tree. The decision tree is built recursively from a training set of trigrams. The TreeTagger also makes use of a full form and a suffix lexicon as well as a prefix lexicon in the case of German. Schmid (1994, 1995) report that the TreeTagger achieves an accuracy of 96.36% on the Penn Treebank data and an accuracy of 97.53% on a small German newspaper corpus of Stuttgarter Zeitung (25,000 tokens in total, 5,000 used for testing). To the best of my knowledge, there is no comparative evaluation of TnT and the TreeTagger on the same German data set. However, on the Swedish Stockholm-Umeå corpus (Ejerhed *et al.*, 1992), the TnT tagger slightly outperforms the TreeTagger at an accuracy of 95.9% versus 95.1%, respectively (Sjöbergh, 2003).

[10]The POS tag sets used in the NEGRA and TIGER corpus annotation is based on the STTS tag set (Schiller *et al.*, 1995). The one used in TIGER is listed in Appendix C. The small differences between that set and the one used in NEGRA are explained in Smith (2003).

Method	Accuracy	Precision	Recall	F-score
Internet				
Baseline	93.95%	-	-	-
TnT_{NEGRA}	98.12%	93.82%	74.35%	82.96
TnT_{TIGER}	98.07%	93.48%	73.31%	82.17
TreeTagger	97.82%	86.52%	76.96%	81.46
Space				
Baseline	96.99%	-	-	-
TnT_{NEGRA}	98.94%	87.09%	76.49%	81.45
TnT_{TIGER}	99.36%	91.38%	87.42%	89.36
TreeTagger	98.08%	62.26%	93.20%	74.65
EU				
Baseline	99.69%	-	-	-
TnT_{NEGRA}	99.70%	48.44%	73.81%	58.49
TnT_{TIGER}	99.78%	60.00%	71.43%	65.22
TreeTagger	99.49%	27.93%	73.81%	40.52

Table 3.17: Task-based evaluation of three POS taggers on the German development data: TnT_{NEGRA} (TnT trained on NEGRA), TnT_{TIGER} (TnT trained on TIGER) and the TreeTagger versus the baseline.

3.5.1.3 Results and Discussion

Table 3.17 lists the performance of the English inclusion classifier when using the various POS taggers and opting for Yahoo in the search engine module. As in the full system, post-processing is applied in all experiments. The results vary per domain. For the internet data, the use of TnT_{NEGRA} results in a slightly higher F-score of 82.96 than when employing TnT_{TIGER} (F=82.17). However, using TnT_{TIGER} yields considerably better results for the other two domains (89.36 and 65.22 versus 81.45 and 58.49, respectively). The TreeTagger results in the worst performance of the English inclusion classifier across all three domains (Internet: F=81.46, Space: F=74.65, EU: F=40.52). Given the fact that the TreeTagger is trained on the least amount of newspaper text (25,000 words) the latter finding is not unexpected. It would be interesting to test the

TreeTagger's performance when trained on either the NEGRA or the TIGER corpus.

As reported in the system description in Section 3.3, the final full English inclusion classifier incorporates TnT_{TIGER} as the POS tagger in the pre-processing step and Yahoo in the search engine module. The decision to use TnT_{TIGER} was made due to the fact that this module results in a drastic performance increase over the TnT_{NEGRA} module for the space travel and EU domains of 7.91 and 6.73 points in F-score, respectively. On the internet data, the TnT_{TIGER} and the TnT_{NEGRA} modules perform very similarly. It can therefore be concluded that a POS tagger trained on a sufficiently large corpus is a vital component of the English inclusion classifier. In the following section, the decision to use Yahoo in the search engine module is motivated.

3.5.2 Task-based Evaluation of Different Search Engines

Tokens that are not found in the German or English lexical database are passed to a back-off search engine module (Section 3.3.4). Such tokens are queried just for German and just for English webpages, a language preference that is offered by most search engines, and classified based on the maximum normalised score of the number of hits returned for each language. This module therefore relies to some extent on the search engine's internal language identification algorithm.

During the initial stages of developing the English inclusion classifier, Google was used in the search engine module (Alex and Grover, 2004; Alex, 2005). The main reasons for opting for Google was that it was the biggest search engine available at the time spanning over 8 billion webpages. It also offers the language preference setting which is essential for determining counts. Moreover, queries can be automated by means of the Google Soap Search API (beta) which allows 1,000 queries per day.[11] During the course of developing the English inclusion classifier, Yahoo, another search engine, also made an API available which allows 5,000 searches per day.[12] In Yahoo, searches can also be restricted to webpages of a particular language. The only differences between the two search engines is their number of indexed webpages. In August 2005, Yahoo announced that it indexes more than 19.2 billion documents[13] which amounts to more than double the number of webpages (8.2 billion) indexed by

[11] `http://www.google.com/apis/`

[12] `http://developer.yahoo.com/`

[13] `http://www.ysearchblog.com/archives/000172.html`

Google. A discussion on the Corpora List in May 2005[14] and a series of studies carried out by Jean Véronis[15] signal that the real number of webpages indexed by a search engine is not necessarily in line with the one that is advertised. Therefore, it is difficult to rely on such quoted figures. Possible artificial inflation of the number of returned hits does however not affect the performance of the English inclusion classifier as long as this inflation occurs consistently for each language-specific query. For example, for Yahoo the estimated English corpus size is 638.9bn tokens whereas that for German is 53.3bn tokens. These estimates illustrate that the English web content is much larger than the German one which is also reflected in the percentages of English and German internet users presented in Figure 2.1, shown in Chapter 2. The ratio between the estimated web corpora for English and German amounts to nearly 12 to 1 in this case. This ratio is similar to those obtained by Grefenstette and Nioche (2000) and Kilgarriff and Grefenstette (2003) (15 to 1 and 11 to 1, respectively) who performed the same estimation using Altavista as the underlying search engine. Before evaluating the use of different search engines with regard to the performance of the English inclusion classifier, the results of a time comparison experiment are reported.[16]

3.5.2.1 Time Comparison Experiment

A comparison of the time required to run the Yahoo module compared to the Google module reveals that the former is considerably faster. Table 3.18 shows the time it takes to estimate the size of the web corpus for three different languages using either Yahoo or Google (Section 3.3.4) which was performed on a 2.4GHz Pentium 4. This estimation involves 16 queries to the search engine API per language. Yahoo clearly outperforms Google by up to 6.1 times.

Web Corpus	German	French	English
Yahoo	6.8s	7.2s	7.6s
Google	35.9s	33.0s	46.4s

Table 3.18: Time required for web corpus estimation using Yahoo and Google.

[14] `http://torvald.aksis.uib.no/corpora/2005-1/0191.html`
[15] `http://aixtal.blogspot.com/2005/08/yahoo-19-billion-pages.html`
[16] All task-based search engine evaluation experiments were conducted in April 2006.

Method	Accuracy	Precision	Recall	F-score
Internet				
Baseline	93.95%	-	-	-
Google	97.96%	92.19%	72.52%	81.21
Yahoo	98.07%	93.48%	73.31%	82.17
Space				
Baseline	96.99%	-	-	-
Google	99.31%	89.83%	87.42%	88.81
Yahoo	99.36%	91.38%	87.42%	89.36
EU				
Baseline	99.69%	-	-	-
Google	99.71%	50.00%	71.43%	58.82
Yahoo	99.78%	60.00%	71.43%	65.22

Table 3.19: Yahoo/Google in a task-based evaluation on the German development data.

3.5.2.2 English Inclusion Classification Experiment

The aim of the following experiment is to determine if the choice of search engine and therefore the language algorithm used by the search engine or the number of indexed webpages has an effect on the performance of the English inclusion classifier when all other parameters are kept the same. Therefore, the classifier is run twice, once with the Google search engine module implementation and once using Yahoo in the search engine module. In order to allow for a clear comparison, the remainder of the system setup is kept the same. As in the full system, post-processing is applied at the end.

Table 3.19 compares the results of both experiments for each domain. There is an improvement in F-scores for all 3 domains when using Yahoo in the search engine module (Internet: +0.96, Space: +0.55, EU: +6.40). The Yahoo module tends to produce a similar recall to the Google module but is more precise. Based on these findings it can be concluded that the choice of search engine affects the performance of the English inclusion classifier. It is difficult to say whether the improvements gained from using Yahoo are due to the fact that it searches more documents than Google alone or due to the search engine's internal language classification. However, calculating

probabilities on a larger corpus tends to yield more robust results in other statistical NLP tasks which certainly explains Yahoo's superiority over Google in this task-based evaluation experiment. Furthermore, opting for Yahoo in the search engine module of the full system considerably speeds up the run time of the English inclusion classifier. The final advantage of using Yahoo is that it allows a larger daily quota of searches (5,000) compared to Google (1,000).

3.6 Machine Learning Experiments

The recognition of foreign inclusions bears great similarity to classification tasks such as NER, for which various machine learning (ML) techniques have proved successful. It is therefore of particular interest to determine the performance of a trained classifier on this task. The following experiments are conducted with a maximum entropy Markov model tagger developed at Standford University which performs well on language-independent NER (Klein *et al.*, 2003) and the identification of gene and protein names (Finkel *et al.*, 2005). In the following, a series of in-domain and cross-domain experiments are discussed, also reported in Alex (2005).[17] The aim is to determine the performance of a supervised ML classifier on unseen data in the domain of the training data as well as on data in a new domain. These results are then compared to those of the annotation-free English inclusion classifier. Moreover, a learning curve created for the ML classifier by training models on increasingly larger training sets illustrates how performance is affected when smaller amounts of training data are available. It also indicates the quantity of labelled training data which is required to achieve a similar performance to that of the English inclusion classifier.

3.6.1 In-domain Experiments

First, several 10-fold cross-validation experiments using different features were conducted on the German development data. They are referred to as in-domain (ID) experiments as the tagger is trained and tested on data from the same domain (see Ta-

[17]The results listed here differ slightly to those reported in Alex (2005) as the best English inclusion classifier has been updated and improved since then. In order to guarantee a fair comparison between the English inclusion classifier and the ML classifier, the cross-validation experiments were rerun on the same data which is POS-tagged with TnT trained on the TIGER corpus.

ble 3.20). In the first experiment (ID1), the tagger's standard feature set is used which includes words, character sub-strings, word shapes, POS tags, abbreviations and NE tags (Finkel *et al.*, 2005). The resulting F-scores are high both for the internet and space travel data (84.74 and 91.29 points) but extremely low for the EU data (13.33 points) due to the sparseness of English inclusions in that data set. ID2 involves the same setup as ID1 but eliminating all features relying on the POS tags. The tagger performs similarly well for the internet and space travel data as for ID1 but improves by 8 points to an F-score of 21.28 for the EU data. This can be attributed to the fact that the POS tagger does not perform with perfect accuracy particularly on data containing foreign inclusions. Training the supervised tagger on POS tag information is therefore not necessarily useful for this task, especially when the data is sparse. Despite the improvement, there is a big discrepancy between the F-score which the ML classifier produces for the EU data and those of the other two data sets.

Table 3.20 compares the best F-scores produced with the tagger's own feature set (ID2) to the best results of the English inclusion classifier presented in this thesis and the baseline. The best English inclusion classifier is the full system combined with consistency checking (Section 3.3.6). For the EU data, the English inclusion classifier performs significantly better than the supervised tagger (χ^2: $df = 1$, $p \leq 0.05$). However, since this data set only contains a small number of English inclusions, this result is not unexpected. It is therefore difficult to draw any meaningful conclusions from these results. For the internet and space travel data sets, which contain many English inclusions, the trained maxent tagger and the English inclusion classifier perform equally well, i.e. without statistical significance between the difference in performance (χ^2: $df = 1$, $p \leq 1$). The fact that the maxent tagger requires hand-annotated training data, however, represents a serious drawback. Conversely, the English inclusion classifier does not rely on annotated data and is therefore much more portable to new domains. Section 3.4.3 shows that it performs well on unseen data in three different domains as well as on entirely new data provided by another research group. Given the necessary lexicons, the English inclusion classifier can easily be run over new text and text in a different language or domain without further cost. The time required to port the classifier to a new language is the focus of attention in the next chapter.

Experiment	Accuracy	Precision	Recall	F-score
Internet				
ID1	98.39%	95.43%	76.23%	84.75
ID2	98.35%	96.38%	74.87%	84.27
ID3	99.23%	95.33%	91.45%	93.35
EIC	98.25%	92.75%	77.37%	84.37
Baseline	93.95%	-	-	-
Space				
ID1	99.51%	99.51%	84.33%	91.29
ID2	99.53%	99.51%	84.54%	91.42
ID3	99.65%	96.30%	91.13%	93.64
EIC	99.45%	89.19%	93.61%	91.35
Baseline	96.99%	-	-	-
EU				
ID1	99.71%	100.00%	7.14%	13.33
ID2	99.73%	100.00%	11.90%	21.28
ID3	99.77%	100.00%	28.57%	44.44
EIC	99.78%	59.26%	76.19%	66.67
Baseline	99.69%	-	-	-

Table 3.20: A series of in-domain (ID) experiments illustrating the performance of a trained ML classifier compared to the English inclusion classifier (EIC) and the baseline.

A further interesting observation is that the ML classifier and the English inclusion classifier perform differently in terms of precision and recall. The tagger is extremely precise but is unable to track all English inclusions. Conversely, the English inclusion classifier is able to identify a larger proportion of English inclusions but some of them by mistake. Therefore, a further experiment (ID3) was conducted, aiming at improving the overall score by combining the behaviour of both systems. ID3 is set up as ID2 but also incorporating the output of the English inclusion classifier as a gazetteer feature. As can be seen in Table 3.20, the tagger's performance increases considerably for all three domains as a result of this additional language feature. The score for the EU data is however still lower than that achieved by the English inclusion classifier itself.

3.6.2 Cross-domain Experiments

In the ID experiments described above, the maxent tagger achieved surprisingly high F-scores for the internet and space travel data, considering the small training sets of around 700 sentences. These high F-scores are mainly attributed to the high precision of the maxent classifier. Although both domains contain many English inclusions, their type-token ratio amounts to 0.29 in the internet and 0.15 in the space travel data (see Table 3.1 in Section 3.2), signalling that English inclusions are frequently repeated in both domains. As a result, the likelihood of the tagger encountering an unknown inclusion in the test data is relatively small which explains high precision scores in the ID experiments.

In order to examine the maxent tagger's performance on data in a new domain containing more unknown inclusions, two cross-domain (CD) experiments were carried out: CD1, training on the internet and testing on the space travel data, and CD2, training on the space travel and testing on the internet data. These two domain pairs were chosen to ensure that both the training and test data contain a relatively large number of English inclusions. Otherwise, the experiments were set up in the same way as experiment ID2 (see Section 3.6.1) using the standard feature set of the maxent tagger minus the POS tag feature. Table 3.21 presents the scores of both CD experiments as well as the percentage of unknown target-type (UTTs). This is the percentage of English types that occur in the test data but not in the training data.

The F-scores for both CD experiments are much lower than those obtained when

	Accuracy	Precision	Recall	F-score	UTT
CD1	98.43%	91.23%	53.61%	67.53	81.9%
EIC	99.45%	89.19%	93.61%	91.35	-
Baseline	96.99%	-	-	-	-
CD2	94.77%	97.10%	13.97%	24.43	93.9%
EIC	98.25%	92.75%	77.37%	84.37	-
Baseline	93.85%	-	-	-	-

Table 3.21: Evaluation scores and percentages of unknown target types (UTT) for two cross-domain (CD) experiments using a maxent tagger compared to the performance of the EIC and the baseline.

training and testing the tagger on documents from the same domain. In experiment CD1, the F-score only amounts to 67.53 points while the percentage of unknown target types in the space travel test data is 81.9%. The F-score is even lower in the second experiment at 24.43 points which can be attributed to the percentage of unknown target types in the internet test data being higher still at 93.9%. These results indicate that the tagger's high performance in the ID experiments is largely due to the fact that the English inclusions in the test data are known, i.e. the tagger learns a lexicon. It is therefore more complex to train a ML classifier to perform well on new data with more and more new anglicisms entering German over time. The amount of unknown tokens will increase constantly unless new annotated training data is added. It can be concluded that the annotation-free English inclusion classifier has a real advantage over any solution that relies on a static set of annotated training data.

3.6.3 Learning Curve

As seen in the previous in- and cross-domain experiments, the statistical ML classifier performs very differently depending on the amount of annotations present in the training data and the domain of that data. In order to get an idea how this classifier performs compared to the English inclusion classifier on a much larger data set, the entire German evaluation data (development and test data for all three domains) was pooled into a large data set containing 145 newspaper articles. As the English inclu-

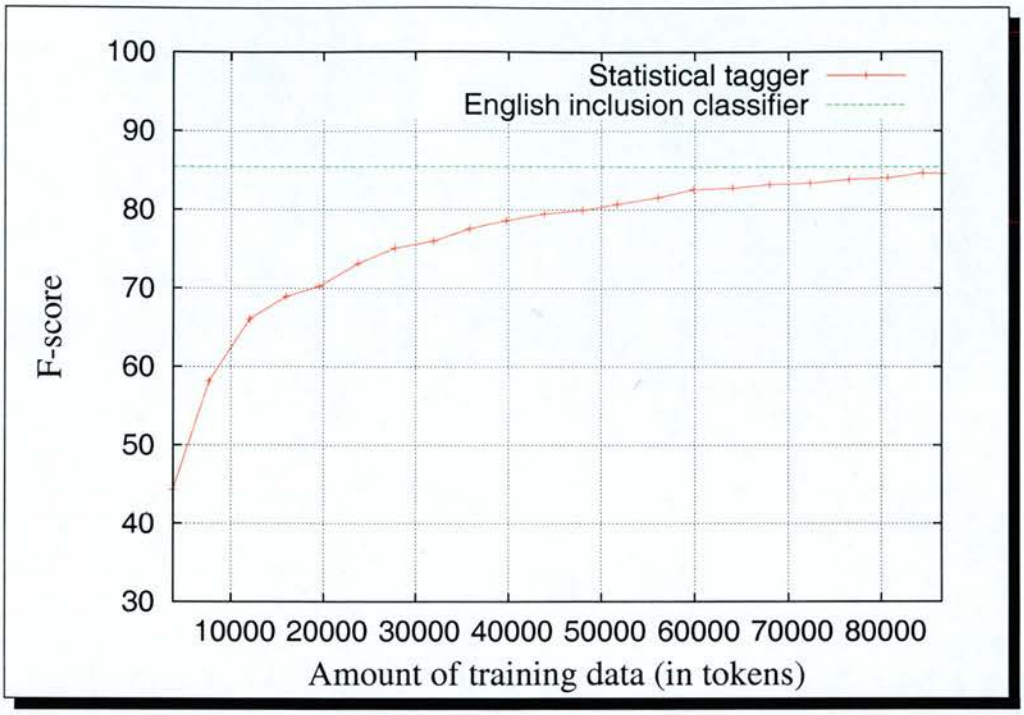


Figure 3.4: Learning curve of a supervised ML classifier versus the performance of the annotation-free English inclusion classifier.

sion classifier does not rely on annotated data, it can be tested and evaluated once for this entire corpus. It yields an overall F-score of 85.43 (see Figure 3.4).

In order to determine the machine learner's performance over the entire data set, and at the same time investigate the effect of the quantity of annotated training data available, a 10-fold cross-validation test was conducted whereby increasingly larger sub-parts of training data are provided when testing on each held out fold. At first, the pooled data is randomised and split into a 90% large training and 10% large test set. This randomisation and split is done on the document level, i.e the training set contains 131 newspaper articles and the test set 14. The training sub-sets are also increased on the document level by batches of 6 newspaper articles at each step. The increasingly larger sub-sets of the training data are then used to train the classifier and subsequently evaluate it on the test set. This procedure is then repeated for each of the 10 held out folds and scores are averaged. Each point in the resulting learning curve presented in Figure 3.4 shows the average F-score of the ML classifier when trained on the selected sub-set of articles and evaluated on the held out set. Average F-scores are plotted

against the average number of tokens in the training data at each step in order to get a better representation of the amount of labelled data involved at each step.

The learning curves presented in Figure 3.4 show that the performance on the ML classifier improves considerably as the amount of training data is increased. The graph shows a rapid growth in F-score which tails off as more data is added. Provided with 100% of labelled training documents (amounting to approximately 86,500 tokens) the ML classifier reaches an F-score of 84.59. The graph shows that the English inclusion classifier has a real advantage over the supervised ML approach which relies on expensive hand-annotated data. A large training set of 86,500 tokens is required to achieve a performance that approximates that of the annotation-free English inclusion classifier. Moreover, the latter system has been shown to perform similarly well on unseen texts in different domains (see Section 3.6.2).

3.7 Chapter Summary

This chapter first described a German newspaper corpus made up of articles on three different topics: internet & telecoms, space travel and European Union. The corpus was annotated in parallel by two different annotators for English inclusions and used for a large number of experiments aimed at English inclusion detection. The corpus analysis showed that, in specific domains, up to 6.4% of the tokens of German newspaper text can be made up of English inclusions. The inter-annotator agreement of identifying English inclusions is very high for a number of metrics, signalling almost perfect agreement and the fact that English inclusion detection is a highly manageable task for humans to carry out.

Subsequently, this chapter presented an annotation-free classifier that exploits linguistic knowledge resources including lexicons and the World Wide Web to detect English inclusions in German text on different domains. Its main advantage is that no annotated, and for that matter unannotated, training data is required. The English inclusion classifier can be successfully applied to new text and domains with little computational cost and extended to new languages as long as lexical resources are available. In the following Chapter, the time and effort involved in extending the system to a new language will be examined. In this chapter, the classifier was examined as whole and in terms of its individual components both on seen and unseen parts of

the German newspaper corpus as well as a test suite of an independent research group.

The evaluation showed that the classifier performs well on unseen data and data in different domains. Its overall performance is approaching the inter-annotator agreement figures which represent an upper bound on the performance that can be expected from an English inclusion classifier. While performing as well as, if not better than, a machine learner which requires a trained model and therefore large amounts of manually annotated data, the output of the English inclusion classifier also increases the performance of the learner when incorporating this information as an additional gazetteer feature. Combining statistical approaches with methods that use linguistic knowledge resources can therefore be advantageous.

The low results obtained in the cross-domain experiments indicate however that the machine learner merely learns a lexicon of the English inclusions encountered in the training data and is unable to classify many unknown inclusions in the test data. The search engine module implemented in the English inclusion classifier is an attempt to overcome this problem as the information on the Web never remains static and at least to some extent reflects language in use. This was reflected in the corpus search module experiments. Moreover, the fact that the English inclusion classifier does not require any manually annotated training data gives it a real advantage over a supervised ML classifier.

Chapter 4

System Extension to a New Language

With increasing globalisation and a rapidly expanding digital society, the influence of English as an international language is growing constantly. As the influx of English vocabulary into other languages is becoming increasingly prevalent, natural language processing systems must be able to deal with this language mixing phenomenon. The previous chapter introduced and evaluated a system that is able to track English inclusions embedded in German text. One criticism that can be made about this English inclusion classifier is that its design is based on a specific language scenario and that it is therefore not language-independent. Therefore, some time was invested in adapting the classifier to a new language in order to investigate the cost involved in doing so. This chapter demonstrates that extending this English inclusion classifier, which was originally designed for German, requires minimal time and effort to adapt to a new language, in this case French. The issue of anglicisms appearing in French was discussed in Section 2.1.3. The existing German system yields high precision, recall and F-scores for identifying English inclusions in unseen data in different domains (Section 3.4.3). In an attempt to carry out similar experiments for a new base language and ascertain the performance for a different language scenario, the system was updated to process French input text as well. The majority of the work presented in this chapter is also reported in Alex (2006).

An indication as to the time necessary to convert each system component of the English inclusion classifier is given in Section 4.1. The French development and test sets created for evaluating the classifier are described in Section 4.2. The extension of individual system modules, which is outlined in Section 4.3, facilitates token level

identification of English inclusions in French text. A series of English inclusion identification experiments on a specially prepared French corpus illustrate the appeal of this system derived from its ease of portability to new languages. Section 4.4 provides a detailed overview of the evaluation experiments and discusses their results which show that the system performs well for both languages and on unseen data in the same domain and language.

4.1 Time Spent on System Extension

The initial English inclusion classifier is designed specifically for German text. The two main aims of extending the system to a new language are: (1) to prove that its underlying concept of English inclusion identification is not specific to one language scenario and (2) to determine the time to do so. In total, it took approximately one person week to convert the core system to French, another Indo-European language with a Latin alphabet. This involved implementing a French tokeniser (1.5 days), incorporating the French TreeTagger (1 day), extending the lexicon module (1.5 days) and converting the search engine module to French (0.2 days).

A subsequent error analysis of the output was performed in order to generate post-processing rules. As the process of analysing errors is essentially difficult to time, a limit of one week was set for this task. This strategy proved beneficial in terms of fine-tuning the system to improve its overall performance (Section 4.4). The actual evaluation of the system requires French data that is manually annotated with English inclusions. Three working days were spent on collecting and annotating a French development and test set of approximately 16,000 tokens each, which are described in more detail in Section 4.2.

A further issue that must be considered when extending the system to a new language is the time required for identifying necessary resources and tools available and familiarising oneself with them. This is evidently dependent on the chosen language. In the case of French, approximately two working days were taken to research and identify the POS tagger TreeTagger and the lexicon Lexique as appropriate resources. If a POS tagger and a lexicon are not available for a particular language scenario, more time and effort would need to be invested to create such resources.

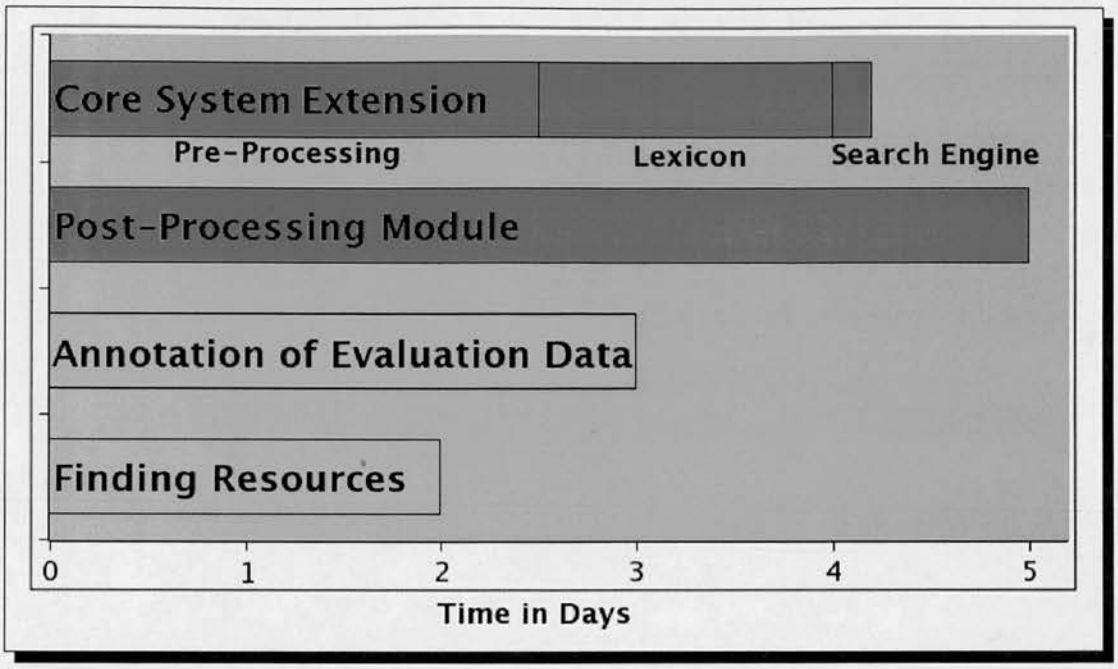


Figure 4.1: Time required for extending system modules as well as data preparation and resource finding tasks.

As the English inclusion classifier is essentially annotation-free, i.e. it does not rely on manually annotated training data, it can be easily run on new data without any further cost. The search engine module then deals with any new vocabulary entering a language over time. This represents a serious advantage over a supervised system that relies on annotated training data. The latter is built on a snapshot of a particular language in use and would need to be adjusted by retraining on additional annotated data as this language evolves over time. It would therefore require much more time and effort to keep up-to-date.

4.2 French Development and Test Data Preparation

In order to evaluate the system performance of classifying English inclusions in French text, a random selection of online articles published by ZDNet France[1], an online magazine reporting on the latest news in the high tech sector, were collected. These articles were published in the period between October 2003 and September 2005 in

[1] http://www.zdnet.fr/

the domain of internet and telecoms (IT). All French articles were manually annotated for English inclusions using an annotation tool based on NXT (Carletta *et al.*, 2003). As with the experiments on German, the data is split into a development set and a test set of approximately 16,000 tokens each. While both sets are of similar domain, date of publishing and content, they do not overlap. The development set served both for the purpose of performance testing of individual system modules for French and error analysis. The test set, however, was treated as unseen and only employed for evaluation in the final run.

Data	Development set					Test set				
Domain: IT	Tokens	%	Types	%	TTR	Tokens	%	Types	%	TTR
French										
Total	16188		3233		0.20	16125		3437		0.21
English	986	6.1	339	10.5	0.34	1089	6.8	367	10.7	0.34
German										
Total	15919		4152		0.26	16219		4404		0.27
English	963	6.0	283	6.8	0.29	1034	6.4	258	5.9	0.25

Table 4.1: Corpus statistics including type-token-ratios (TTRs) of the French development and test sets compared to the German data.

Table 4.1 lists the total number of tokens and types plus the number of English inclusions in the French development and test sets. The corresponding statistics of the German data are reproduced to facilitate a comparison. The French data sets have similar characteristics, particularly regarding their type-token-ratios (TTRs) for each entire set (0.20 versus 0.21) and for the English inclusions alone (0.34 each). The French test set contains slightly more English inclusions (+0.7%) than the development set. Comparing these figures with those of the previously annotated German IT data sets shows that the proportion of English tokens in this domain is extremely similarly at approximately 6%. However, the percentage of English types varies to some extent both for the development and test sets. They only amount to 6.8% and 5.9% in the German data, compared to 10.5% and 10.7% in the French data. Moreover, the TTRs of English inclusions are 0.5 and 0.9 points higher in the French data sets, signalling

that they are less repetitive than those contained in the German articles. However, overall TTRs are 0.6 points lower for French than for German which means that the remaining vocabulary in the French articles is somewhat less heterogeneous than in the German data. This is due to the nature of the two languages. In German, lexical variety is higher due to compounding and case inflection for articles, adjectives and nouns.

French internet data			
Token	f	Token	f
e	49	web	21
Google	44	spam	18
internet	35	mail	18
Microsoft	33	Firefox	16
mails	32	Windows	14

Table 4.2: Ten most frequent (f) English inclusions in the French data.

The ten most frequent English inclusions found in the French data set are listed in Table 4.2. The majority of them are either common or proper nouns. The most frequent English token is actually the single-character token *e* as occurring in *e-mail*. Less frequent pseudo-anglicisms like *le forcing* in the sense of putting pressure on someone or *le parking* referring to a car park are also contained in the French data. Although such lexical items are either non-existent or scarcely used by native English speakers, they are also annotated as English inclusions.

4.3 System Module Conversion to French

The extended system architecture of the English inclusion classifier consists of several pre-processing steps, followed by a lexicon module, a search engine module and a post-processing module. Converting the search engine module to a new language required little computational cost and time. Conversely, the pre- and post-processing as well as the lexicon modules necessitated some language knowledge resources or

tools and therefore demanded more time and effort to be customised for French. The core system was adapted in approximately one person week in total (Section 4.1). Figure 4.2 illustrates the system architecture after extending it to French. Note that the system now involves an additional document-based language identification step after pre-processing in which the base language of the document is determined by TextCat (Cavnar and Trenkle, 1994). TextCat, a traditional language identification tool, performs well on identifying the language of sentences and larger passages. This enables running the English inclusion classifier over either German or French text without having to specify the base language of the text manually. The base-language-specific classifier components are therefore initiated purely based on TextCat's language identification. For both the German and French newspaper articles, TextCat is always able to identify the language correctly.

4.3.1 Pre-processing Module

The pre-processing module involves tokenisation and POS tagging (cf. Section 3.3.2). First, the German tokeniser was adapted to French and a French part-of-speech (POS) tagger was integrated into the system. The French tokeniser consists of two rule-based tokenisation grammars. In the same way as the German version, it not only identifies tokens surrounded by white space and punctuation but also resolves typical abbreviations, numerals and URLs. Both grammars are applied by means of improved upgrades of the XML tools described in Thompson *et al.* (1997) and Grover *et al.* (2000). These tools process an XML input stream and rewrite it on the basis of the rules provided. The French TreeTagger (see Section 3.5.1.2) is used for POS tagging. It is freely available for research and is also trained for a number of other languages, including German and English. The TreeTagger functions on the basis of binary decision trees trained on a French corpus of 35,448 word forms and yields a tagging accuracy of over 94% on an evaluation data set comprising of 10,000 word forms (Stein and Schmidt, 1995).[2]

[2]While trained models are available online, the tagged data set that was used to train and evaluate the French TreeTagger model is not part of the distribution. Otherwise, the data could have been used to train TnT, as that tagger resulted in a better performance of the English inclusion classifier on the German development data (see Section 3.5.1).

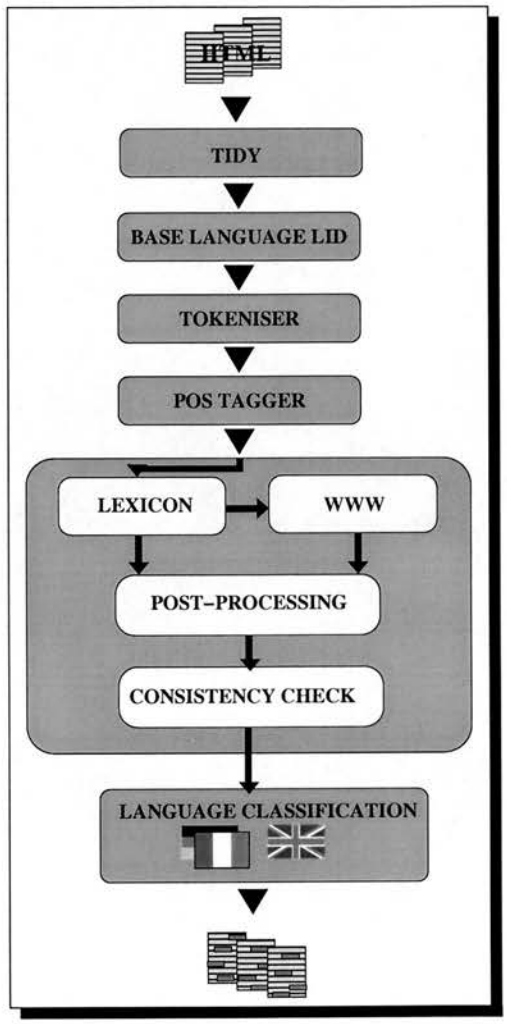


Figure 4.2: Extended system architecture.

4.3.2 Lexicon Module

The lexicon module (cf. Section 3.3.3) performs an initial language classification run based on a case-insensitive double lookup procedure using two lexicons: one for the base-language and one for the language of the inclusions. For French, the system queries Lexique, a lexical database which contains 128,919 word forms representing 54,158 lemmas (New *et al.*, 2004). It is derived from 487 texts (31 million words) published between 1950 and 2000. In order to detect common English inclusions, the system searches the English database of CELEX holding 51,728 lemmas and their 365,530 corresponding word forms (Celex, 1993) . The lexicon module was adapted to French by exploiting distinctive characteristics of French orthography. For example, words containing diacritic characters typical for French are automatically excluded from being considered as English inclusions.[3]

4.3.3 Search Engine Module

Tokens which are not clearly identified by the lexicon module as English are passed to the back-off search engine module. As explained in Section 3.3.4, the search engine module performs language classification based on the maximum normalised score of the number of hits returned for two searches per token, one for each language L:

$$rf_{C_{web}(L)}(t) \tag{4.1}$$

Extending the search engine module to French merely involved adjusting the language preferences in the search engine API and incorporating the relative frequencies of representative French tokens in a standard French corpus for estimating the size of the French Web corpus (Grefenstette and Nioche, 2000). The search engine Yahoo was used instead of Google as it outperformed the latter in a task-based evaluation for German (Section 3.5.2) and also allows a larger number of automatic queries per day.

[3]Note that English also contains words with diacritic characters, e.g. *née* or *naïve*. However, they tend to be loan words from French or other languages. When appearing in French text, they would be either part of a French word or a word from a language other than English.

4.3.4 Post-processing Module

The final system component which required adapting to French is the post-processing module. It is designed to resolve language classification ambiguities and classify some single-character tokens. As for German, some time was invested in analysing the core system output of the French development data in order to generate these post-processing rules. The individual contribution of each of the following rules on the system performance on the French development data is discussed in Section 4.4.

The most general rules are designed to disambiguate one-character tokens and interlingual homographs. They are flagged as English if followed by a hyphen and an English token (*e-mail* or *joint-venture*). Furthermore, typical English function words are flagged as English, including prepositions, pronouns, conjunctions and articles, as these belong to a closed class and are easily recognisable. This also avoids having to extend the core system to these categories which not only prevents some output errors but also improves the performance of the POS tagger which is often unable to process foreign function words correctly. In the post-processing step, their POS tags are therefore corrected. Any words in the closed class of English function words that are ambiguous with respect to their language such as *an* (in French *year*) or *but* (in French *goal*) are only flagged as English inclusions if their surrounding context is already classified as English by the system. Similarly, the possessive marker *'s* is flagged as English if it is preceded by an English token. Moreover, several rules are designed to automatically deal with names of currencies (e.g. *Euro*) and units of measurement (e.g. *Km*). Such instances are prevented from being identified as English inclusions. Similarly to the German post-processing module, abbreviation extraction (Schwartz and Hearst, 2003) is applied in order to relate language information between abbreviations or acronyms and their definitions. Subsequently, post-processing is applied to guarantee that each pair and earlier plus later mentions of either the definition or the abbreviation/acronym are assigned the same language tag within a document.

When analysing the errors which the system made in the development data, it was also observed that foreign person names (e.g. *Graham Cluley*) are frequently identified as English inclusions. In the experiments described in this thesis, the system is evaluated on identifying English inclusions. These are defined as any English words in the text except for person and location names. Therefore, the evaluation data does

not contain annotations of foreign, or specifically English person names in the gold standard. In order to improve the performance of recognising real English inclusions, further post-processing rules were implemented to distinguish between the latter and English person names that are incorrectly classified as English inclusions. The aim is to increase precision without reducing recall. Based on an error analysis on the development data, patterns signalling person names in French text, e.g. "Mme X" or "X, directeur", were generated to distinguish such instances from English inclusions. It should be noted, however, that for a potential task-based evaluation of the English inclusion classifier the language information provided for person names could prove beneficial, for example for TTS synthesis to generate correct pronunciations.

Combining all post-processing rules allows the English inclusion classifier to perform with a balanced F-score of 87.68 (P=91.60% and R=84.08%) on the French development data. This represents an overall performance improvement of 16.09 points in F-score, 8.69% in precision and 21.10% in recall over the core system (see Table 4.3). These results show that post-processing is mainly aimed at identifying false negatives, i.e. English inclusions which are missed by the core system. As is the case for the German data, the precision of the core system is already relatively high.

4.4 French System Evaluation

This section evaluates the performance of the French system, compared to the German one, when testing it on unseen data from a similar domain, i.e. internet & telecoms related data. Furthermore, it presents some additional results illustrating the improvement gained from the various post-processing rules and from added document consistency checking.

4.4.1 Evaluation on Test and Development Data

Table 4.3 shows the results of the core and full French and German systems on the development and test data versus the baseline. The full system includes post-processing as well as document consistency checking. The baseline accuracies are determined by assuming that the system found none of the English inclusions in the data and believes that it is entirely written in either French or German, respectively. As precision,

	Test set				Development set			
Method	A	P	R	F	A	P	R	F
French data								
BL	93.25%	-	-	-	93.91%	-	-	-
CS	96.59%	82.07%	69.33%	75.16	96.74%	82.91%	62.98%	71.59
FS	98.08%	88.35%	84.30%	86.28	98.49%	91.39%	85.09%	88.13
German data								
BL	93.62%	-	-	-	93.95%	-	-	-
CS	97.47%	90.60%	66.32%	76.58	97.15%	87.28%	67.70%	76.25
FS	98.13%	91.58%	78.92%	84.78	98.25%	92.75%	77.37%	84.37

Table 4.3: Evaluation of the core and full systems (CS and FS) on the French and German development and unseen test sets versus the baseline (BL).

recall and F-score are determined in relation to the English tokens in the gold standard (see Appendix A.1), they are essentially zero for the baseline. For this reason, only the accuracy baseline scores are reported, as was done in all previous evaluation experiments.

The German core system (without post-processing and consistency checking) performs similarly on both the development set and the test set at approximately 76 points in F-score. The French core system actually performs almost 4 points better in F-score on the test set (F=75.16) than on the development set (F=71.59). This means that the core systems do not undergenerate on new data in the same domain and language. Comparing the results of the core systems across languages shows that they perform relatively similarly in F-score but vary slightly in terms of precision and recall. These differences can be attributed to some system internal differences resulting from language-specific characteristics or pre-processing tools. For example, 13.7% of all tokens in the French development set contain diacritics compared to only 7.8% of all tokens in the German development set. As information about diacritics is exploited in the lexicon module for both languages, the French system is expected to perform better at that stage.

A further core system difference lies in the POS tag sets for the two languages. The German system makes use of TnT (Brants, 2000b) trained on the TIGER Treebank (Brants *et al.*, 2002) to assign POS tags. Earlier experiments showed that this

POS tagger yields the best results for a set of German data sets in different domains (Section 3.5.1). TnT assigns STTS tags to German text (Schiller *et al.*, 1995). The English inclusion classifier is set up to process any token in the German data with the tag: NN (common noun), NE (proper noun), ADJA or ADJD (attributive and adverbial or predicatively used adjectives) as well as FM (foreign material). The French data, on the other hand, is tagged with the TreeTagger whose POS tag set differs to STTS. Although it also differentiates between common nouns (NOM) and proper names (NAM), it only has one tag for adjectives (ADJ). Moreover, the French tag set contains an additional abbreviation tag (ABR). It does not, however, contain a separate tag for foreign material. Despite the fact that TnT is not very accurate in identifying foreign material in the German data, this additional information is likely to have a positive effect on the overall performance of the German system.

The full system scores show that post-processing and document consistency checking improve the overall system performance considerably for both languages. For French, the individual post-processing rules are evaluated in more detail in the next section. Overall, Table 4.3 shows that the improvements are relatively similar on both the development set and the test set for each language. The full French system performs only 1.85 points lower in F-score on the test set (86.28 points) compared to the development set (88.13 points). The full system scores for German are also very similar. Within each language, the classifier therefore produces consistent results.

Overall, the full French system performs slightly better than the German one. Table 4.3 illustrates that this difference is mainly due to the larger gap between recall and precision for the full German system. Even though the full German system performs better in precision than the French one, its recall is much lower, causing the overall F-score to drop. This discrepancy is due to language-specific post-processing differences as post-processing rules are generated on the basis of error analysis on the development data. However, comparing the results of the two systems is not entirely straightforward because they are not completely identical in parts of their components and the data sets are inherently not identical. Despite these differences, the fact that both systems yield high F-scores demonstrates that the underlying concept of identifying English inclusions in text can be applied to different language scenarios, at least those with Latin alphabets.

4.4.2 Evaluation of the Post-Processing Module

Table 4.4 presents lesion studies showing the individual contribution of different types of post-processing rules to the overall performance of the full French system on the development data. In this case, the full system does not include document consistency checking. A detailed description of the post-processing module design is given in Section 4.3.4.

The results show that the biggest improvement is due to the post-processing of single-character tokens which are not classified by the core system. Switching off this type of post-processing leads the full system to perform 7.16 points lower in F-score. The second largest improvement in F-score is achieved by the post-processing rules dealing with ambiguous words, i.e. those that are classified as either French or English by the core system. Identifying the language of such tokens based on the language of their surrounding context helps to improve the overall performance by 5.73 points in F-score. Comparing the results of each of the lesion study experiments to the results of the full system, where all post-processing rules are applied, also shows that most post-processing rules are designed to improve recall.

The only post-processing rule implemented to improve precision without deteriorating recall is that for person names. It results in a smaller but nevertheless statistically significant increase (χ^2: $df = 1$, $p \leq 0.001$) of 2.39 points in F-score. Although all remaining post-processing rules do not yield statistically significant performance improvements, none of the post-processing rules lead to a decrease in F-score as is illustrated in the last column. In the final run of the full French system on the test data, the post-processing module results in a large performance increase of 11.13 points in F-score. Therefore, it can be concluded that the post-processing is designed well enough to apply to new data in the same domain and language.

4.4.3 Consistency Checking

In order to guarantee consistent language classification within each document, additional consistency checking was implemented in the French system. This second classification run functions the same way as that implemented in the German system, described in Section 3.3.6. Tokens that are identified as English by the full system after

Post-processing	Accuracy	Precision	Recall	F-score	Δ F
None	96.74%	82.91%	62.98%	71.59	16.57
Single letters	97.75	90.51%	72.52%	80.52	7.16
Ambiguous words	97.83	91.60%	74.14%	81.95	5.73
Person names	98.12%	86.53%	84.08%	85.29	2.39
Function words	98.21%	91.36%	81.54%	86.17	1.51
Currencies etc.	98.30%	91.08%	81.85%	86.22	1.46
Abbreviations	98.39%	90.87%	83.77%	87.18	0.50
Full System - CC	98.45%	91.60%	84.08%	87.68	-

Table 4.4: Evaluation of the post-processing module with one type of post-processing removed at a time on the French development data. Δ F represents the change in F-score compared to the full English inclusion classifier without consistency checking (CC).

post-processing are added to a gazetteer. This gazetteer is then checked on the fly to assure that tokens that were not already previously tagged by the system are classified correctly as well. Consistency checking is therefore mainly aimed at identifying English inclusions which the POS tagger did not tag correctly. For example, the word *Google* was once incorrectly tagged as a present tense verb (`VER:pres`) and could therefore not be classified by the system initially. However, since the same token was also listed in the on-the-fly gazetteer which was generated for the particular document it occurred in, consistency checking resulted in the correct classification.

Table 4.5 presents the performance of the full French and German systems with optional consistency checking on both the development and test data. The results show that consistency checking does not have the same effect on the French as it does on the German data. It only yields a small improvement in F-score of 0.45 points on the French development data but no improvement on the French test data. One reason for this discrepancy between languages could be the POS tagging of English inclusions. While English inclusions in the German development data are assigned on average 1.2 POS tags by TnT, the TreeTagger tags the English inclusions in the French development data only with 1.1 different POS tags. The latter is therefore slightly more consistent. The second reason is that English inclusions are repeated less often in the French data than in the German which is demonstrated in their TTRs (0.34 in French

development and test sets versus 0.29 and 0.25 in German development and test sets, see Table 4.1). This means that the classifier is already less likely to miss inclusions which minimises the effect of consistency checking for French.

	Test set				Development set			
Method	Acc	P	R	F	Acc	P	R	F
French data								
FS	98.10%	88.59%	84.11%	86.29	98.45%	91.60%	84.08%	87.68
FS+CC	98.08%	88.35%	84.30%	86.28	98.49%	91.39%	85.09%	88.13
German data								
FS	97.93%	92.13%	75.82%	83.18	98.07%	93.48%	73.31%	82.17
FS+CC	98.13%	91.58%	78.92%	84.78	98.25%	92.75%	77.37%	84.37

Table 4.5: Evaluation of the full system (FS) with optional consistency checking (CC).

4.5 Chapter Summary

This chapter described how the English inclusion classifier was successfully converted to a new language, French. The extended system is able to process either German or French text for identifying English inclusions. The system is a pipeline made up of several modules, including pre-processing, a lexicon, a search-engine, a post-processing and a document consistency checking module. The extension of the core system was carried out in only one person week and resulted in a system performance of 71.59 points in F-score on the French development data. A further week was spent on implementing the post-processing module which boosted the F-score to 87.68 points. A third week was required to select external language resources plus collect and annotate French evaluation data in the domain of internet and telecoms.

The performance drop between the French development set and the unseen test sets is relatively small (1.85 in F-score) which means that the system does not seriously over- or undergenerate for this domain but results in an equally high performance on new data. This chapter also demonstrated that the English inclusion classifier is easily extendable to a new language in a relative short period of time and without having to

rely on expensive manually annotated training data. Therefore non-recoverable engineering costs for extending and updating the classifier are kept to a minimum. Not only can the system be easily applied to new data from the same domain and language without a serious performance decrease, it can also be extended to a new language and produce similarly high scores. The performance could possibly be even better for languages with the same script that are less closely related than French and English or German and English.

The English inclusion classifier described in this and the previous chapter is designed particularly for languages composed of tokens separated by white space and punctuation and with Latin-based scripts. A system that tracks English inclusions occurring in languages with non-Latin based scripts necessitates a different setup as the inclusions tend to be transcribed in the alphabet of the base language of the text (e.g. in Russian). The English inclusion classifier is also not designed to deal with languages where words are not separated by white space. An entirely different approach would be required for such a scenario.

Chapter 5

Parsing English Inclusions

The status of English as a global language means that English words and phrases are frequently borrowed by other languages, especially in domains such as science and technology, commerce, advertising and current affairs. This is an instance of *language mixing*, whereby inclusions from other languages appear in an otherwise monolingual text. While the processing of foreign inclusions has received some attention in the TTS literature (see Chapter 6), the natural language processing (NLP) community has paid little attention both to the problem of inclusion detection, and to potential applications thereof. Also, the extent to which inclusions pose a problem to existing NLP methods has not been investigated, a challenge addressed in this chapter.[1]

The main focus is on the impact which English inclusions have on the parsing of German text. Anglicisms and other borrowings from English form by far the most frequent foreign inclusions in German. In specific domains, up to 6.4% of the tokens of a German newspaper text can be made up of English inclusions. Even in regular newspaper text processed by many NLP applications, English inclusions can be found in up to 7.4% of all sentences (see Table 3.1 and 5.2 for both figures).

Virtually all existing NLP algorithms assume that the input is monolingual and does not contain foreign inclusions. It is possible that this is a safe assumption, and inclusions can be dealt with accurately by existing methods, without resorting to specialised mechanisms. The alternative hypothesis, however, seems more plausible: foreign inclusions pose a problem for existing approaches and sentences containing them are processed less accurately. A parser, for example, is likely to have difficulties with pro-

[1]The content of the first part of this chapter is also published in Alex *et al.* (2007).

cessing inclusions. Most of the time, they are unknown words and, as they originate from another language, standard methods for unknown word guessing (suffix stripping, etc.) are unlikely to be successful. Furthermore, the fact that inclusions are often multi-word expressions (e.g., named entities or code-switches) means that simply part-of-speech (POS) tagging them accurately is not sufficient: the parser positing a phrase boundary within an inclusion is likely to severely decrease accuracy.

After a brief summary of related work in Section 5.1, this chapter then describes an extrinsic evaluation of this classifier for parsing. It is shown that recognising and dealing with English inclusions via a special annotation label improves the accuracy of parsing. In particular, this chapter demonstrates that detecting English inclusions in German text improves the performance of two German parsers, a treebank-induced parser as well as a parser based on a hand-crafted grammar (Sections 5.3 and 5.4). Crucially, the former parser requires modifications of its underlying grammar to deal with the inclusions, the latter's grammar is already designed to deal with multi-word expressions signalled in the input. Both parsers and necessary modifications are described in detail in Sections 5.3.1 and 5.4.1. The data used for all the parsing experiments is described in 5.2.

5.1 Related Work

Previous work on inclusion detection exists in the TTS literature (Pfister and Romsdorfer, 2003; Farrugia, 2005; Marcadet *et al.*, 2005), which is reviewed in detail in Sections 2.2.1.1 and 2.2.1.2. Here, the aim is to design a system that recognises foreign inclusions on the word and sentence level and functions as the front-end to a polyglot TTS synthesiser. Similar initial efforts have been undertaken in the field of lexicography where the importance of recognising anglicisms from the perspective of lexicographers responsible for updating lexicons and dictionaries has been acknowledged (Andersen, 2005) (see also Section 2.2.1.4). In the context of parsing, however, there has been little focus on this issue. Although Forst and Kaplan (2006) have stated the need for dealing with foreign inclusions in parsing as they are detrimental to a parser's performance, they do not substantiate this claim using numeric results.

Previous work reported in this thesis have focused on devising a classifier that detects anglicisms and other English inclusions in text written in other languages, namely

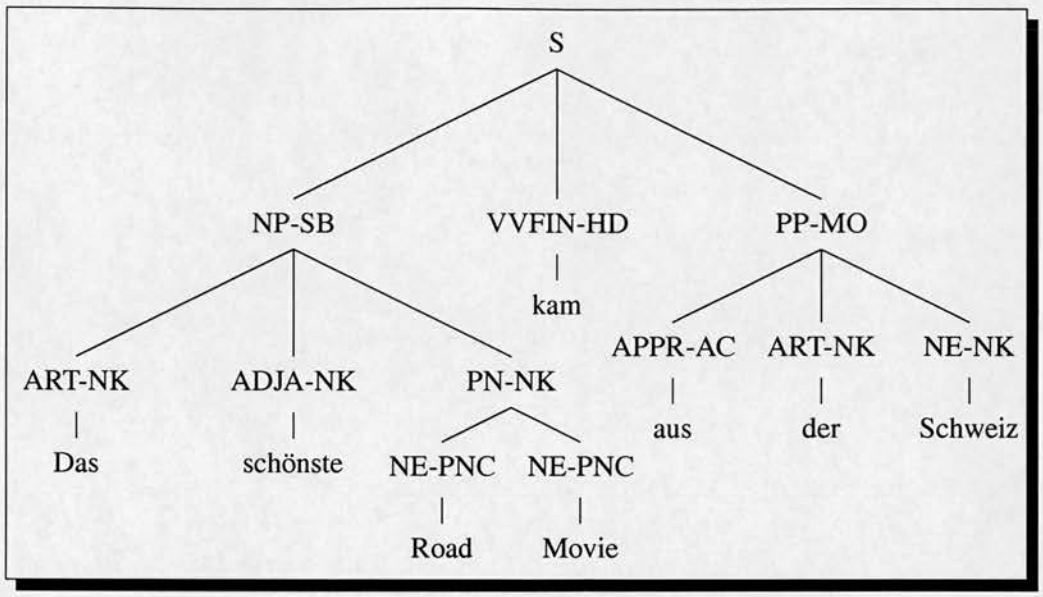


Figure 5.1: Example parse tree of a German TIGER sentence containing an English inclusion. Translation: *The nicest road movie came from Switzerland.*

German and French. In Chapter 3, it has been shown that the frequency of English inclusions varies considerably depending on the domain of a text collection but that the classifier is able to detect them equally well with an F-score approaching 85 for each domain.

5.2 Data Preparation

The experiments described in this chapter involve applying the English inclusion classifier to the TIGER treebank (Brants *et al.*, 2002)[2], a syntactically annotated corpus consisting of 40,020 sentences of German newspaper text, and evaluating it extrinsically on a standard NLP task, namely parsing. The aim is to investigate the occurrence of English inclusions in general newspaper text and to examine if the detection of English inclusions can improve parsing performance. The English inclusion classifier was run once over the entire TIGER corpus. In total, the system detected English

[2]All the following parsing experiments are conducted on TIGER data (Release 1). Some of them contain additional language knowledge output by the English inclusion classifier. The pre-processing module of the classifier hereby always involves POS tagging with the TnT tagger trained on the NEGRA corpus (TnT_{NEGRA}, see Section 3.5.1.1) and not the TIGER corpus.

inclusions in 2,948 of 40,020 sentences (7.4%), 596 of which contained at least one multi-word inclusion. This sub-set of 596 sentences is the focus of the work reported in the remainder of this chapter, and will be referred to as the inclusion set.

A gold standard parse tree for a sentence containing a typical multi-word English inclusion is shown in Figure 5.1. It can be seen that the tree is relatively flat, which is a common characteristic of TIGER treebank annotations (Brants *et al.*, 2002). The non-terminal nodes of the tree represent a combination of both the phrase categories (e.g. noun phrase (NP)) and the grammatical functions (e.g. subject (SB)). In the example sentence, the English inclusion is contained in a proper noun (PN) phrase with a grammatical function of type noun kernel element (NK). Each terminal node is POS-tagged as a named entity (NE) with the grammatical function of type proper noun component (PNC).

5.2.1 Data Sets

For the following experiments, two different data sets are used:

1. the inclusion set, i.e. the sentences containing multi-word English inclusions recognised by the inclusion classifier, and

2. a stratified sample of sentences randomly extracted from the TIGER corpus, with strata for different sentence lengths.

For the stratified sample, the strata were chosen so that the sentence length distribution of the random set matches that of the inclusion set. The average sentence length of this random set and the inclusion set is therefore the same at 28.4 tokens. This type of sampling is necessary as parsing performance is correlated with sentence length. The inclusion set has a higher average sentence length than a completely random sample of sentences extracted from the TIGER corpus (as is displayed in Figure 5.2) which only amounts to 17.6 tokens. Both the inclusion set and the stratified random set consist of 596 sentences and do not overlap. They are used in the experiments with the treebank-induced parser and the hand-crafted grammar described in Sections 5.3 and 5.4, respectively.

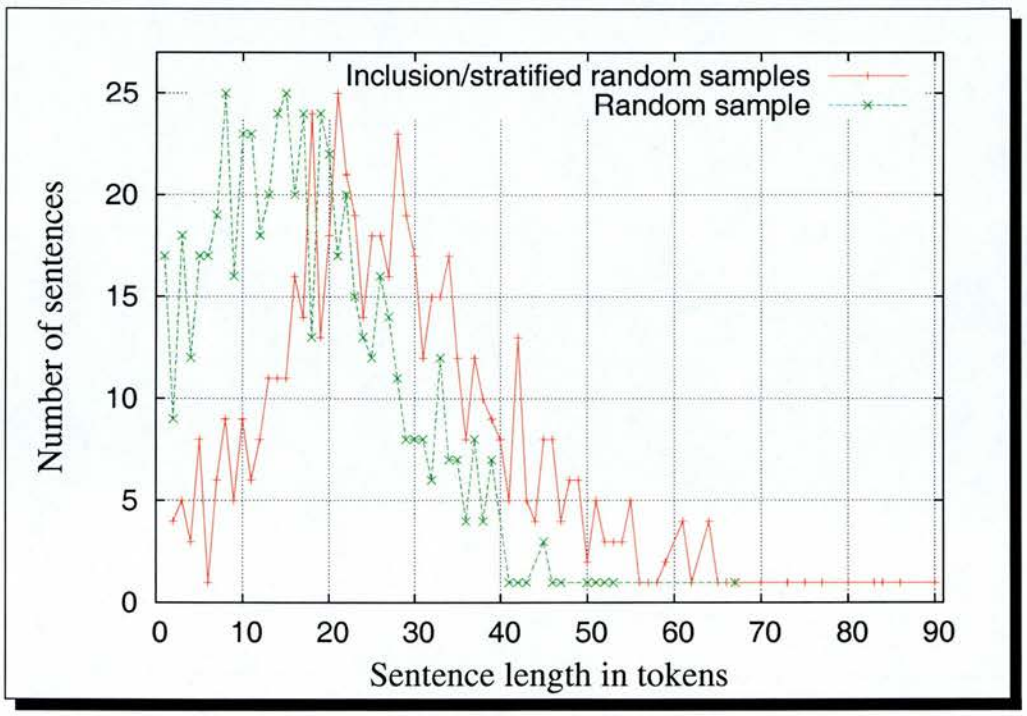


Figure 5.2: Sentence length distribution of the inclusion set and a completely random TIGER data set, containing 596 sentences each.

5.3 Treebank-induced Parsing Experiments

5.3.1 Parser

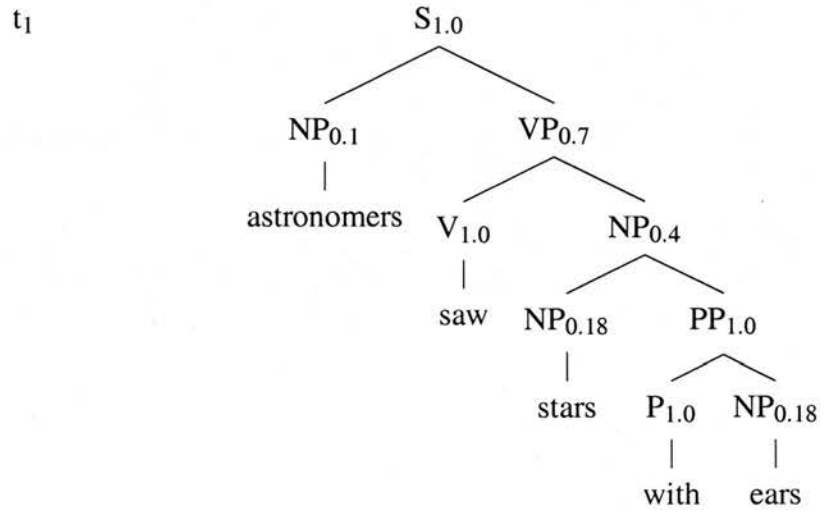
The first set of parsing experiments were performed with a treebank-induced parser trained on the TIGER corpus which returns both phrase categories and grammatical functions (Dubey, 2005b). Following Klein and Manning (2003), the parser uses an unlexicalised PCFG, a probabilistic context-free grammar (Booth and Thompson, 1973). A PCFG consists of a set of terminals, non-terminals, a start symbol, a set of rules and a corresponding set of associated rule probabilities. The overall probability of a given tree is determined by multiplying all probabilities of local rules. Context-free means that the probabilities of sub-trees do not depend on words which are not dominated by the sub-tree (Manning and Schütze, 2001). An example parse tree and its probabilities for a toy PCFG are displayed in Figure 5.3.

Dubey's parser determines the local rule probabilities from a collection of correctly parsed example sentences.[3] This means that the probabilistic full parsing model is induced by training on a syntactically annotated corpus, called a treebank. For example, Dubey (2005a,b) reports parsing performance for models trained on the German NEGRA treebank (Skut *et al.*, 1998). The main characteristic of the parser is that it is unlexicalised, which, in contrast to English, Dubey and Keller (2003) found to outperform lexicalised parsing algorithms in German. A convenient property of the parser is that it can be trained on a new treebank. Furthermore, Dubey's parser relies on automatic treebank transformations to increase parsing accuracy. Crucially, these transformations make use of TIGER's grammatical functions to relay pertinent lexical information from lexical elements up into the tree. The principal reason for applying these treebank transformations, also referred to as grammatical function reannotations, is to overcome data sparseness. For example, a coordinated accusative noun phrase rule (see Figure 5.4(a)) fails to explicitly state that its coordinate sisters are accusative objects (`OA`) but only signifies that they are part of a conjunction (`CJ`). Therefore, a transformation is applied to replace the original rule with the one shown in Figure 5.4(b) which makes the case information explicit in the pre-terminal nodes.

Based upon an evaluation on the NEGRA treebank, using a 90%-5%-5% training-

[3] Dubey's software allows automatic POS tagging as part of the parsing process as his parser learns grammar rules that extend to POS tags from the training data.

S → NP VP	1.0	NP → NP PP	0.4
PP → PP NP	1.0	NP → *astronomers*	0.1
VP → V NP	0.7	NP → *ears*	0.18
VP → VP PP	0.3	NP → *saw*	0.04
P → *with*	1.0	NP → *stars*	0.18
V → *saw*	1.0	NP → *telescopes*	0.1



$P(t_1) = 0.0009072$

Figure 5.3: Parse tree with local rule probabilities shown for an English example sentence based on a simple PCFG (Manning and Schütze, 2001).

CNP-OA→NP-CJ KON-CD NP-CJ

(a) Original coordinated accusative NP rule.

CNP-OA→NP-OA KON-CD NP-OA

(b) Transformed coordinated accusative NP rule.

Figure 5.4: Tree transformation for a coordinated noun phrase rule.

development-test split, the parser performs with an accuracy of 73.1 F-score on labelled brackets with a coverage of 99.1% (Dubey, 2005b). Dubey (2005b) has found that, without affecting coverage, the transformations improve parsing performance by 4 points in F-score over the baseline grammatical function parser which yields an F-score of 69.1 on the NEGRA test set.

In addition to the treebank re-annotation, the parser also makes use of suffix analysis, however, beam search or smoothing are not employed. Both beam search and smoothing lead the model to perform better but result in a decrease in coverage and an increase in parsing time by up to 10 times, respectively (Dubey, 2005a). Dubey's figures are derived on a test set limited to sentences containing 40 tokens or less. In the data sets used in the experiment that are presented in this chapter, however, sentence length is not limited. Moreover, the average sentence length of these test sets is considerably higher (28.4 tokens) than that of the NEGRA test set (17.24 tokens). Consequently, a slightly lower performance and/or coverage is anticipated, even though the type and domain as well as the annotation of both the NEGRA and the TIGER treebanks are very similar. The minor annotation differences that do exist between NEGRA and TIGER are explained by Brants *et al.* (2002).

5.3.2 Parser Modifications

Several variations of the parser are tested: (1) the baseline parser, (2) the perfect tagging model, (3) the word-by-word model and (4) the inclusion entity model. The baseline parser does not treat foreign inclusions in any special way, i.e. the parser attempts to guess the POS tag of each inclusion token using the same suffix analysis as for rare or unseen German words. The additional versions of the parser are inspired by

the hypothesis that inclusions make parsing difficult, and this difficulty arises primarily because the parser cannot detect inclusions. Therefore, an anticipated upper bound is to give the parser perfect tagging information. Two further versions interface with the English inclusion classifier and treat words marked as inclusions differently from native words. The first version does so on a word-by-word basis. Conversely, the second version, the inclusion entity approach, attempts to group inclusions even if a grouping is not posited by phrase structure rules. Each version is now described in detail.

5.3.2.1 Perfect Tagging Model

This model involves allowing the parser to make use of perfect tagging information for all tokens given in the pre-terminal nodes. In the TIGER annotation, pre-terminals include not only POS tags and but also grammatical function labels. For example, rather than a pre-terminal node having the category `PRELS` (personal pronoun), it is given the category `PRELS-OA` (accusative personal pronoun) in the gold standard annotation. When given the POS tags along with the grammatical functions, the perfect tagging parser may unfairly disambiguate more syntactic information than when simply provided with perfect POS tags alone. Therefore, to make this model more realistic, the parser is required to guess the grammatical functions itself, allowing it to, for example, mistakenly tag an accusative personal pronoun as a nominative, dative or genitive one. This setup gives the parser access to information about the gold standard POS tags of English inclusions along with those of all other words, but does not offer any additional hints about the syntactic structure of the sentence as a whole.

5.3.2.2 Word-by-word Model

The two remaining models both take advantage of information acquired from the English inclusion classifier. To interface the classifier with the parser, each inclusion is simply marked with a special `FOM` (foreign material) tag. The word-by-word parser attempts to guess POS tags itself, much like the baseline. However, whenever it encounters a `FOM` tag, it restricts itself to the set of POS tags observed for inclusions during training (the tags listed in Table 5.1). When a `FOM` is detected, these and only these POS tags are guessed; all other aspects of the parser remain the same.

POS-tag	NE	FM	NN	KON	CARD	ADJD	APPR
Count	1185	512	44	8	8	1	1

Table 5.1: POS tags of English inclusions.

5.3.2.3 Inclusion Entity Model

The word-by-word parser fails to take advantage of one important trend in the data: that foreign inclusion tokens tend to be adjacent and these adjacent words usually refer to the same entity. There is nothing stopping the word-by-word parser from positing a constituent boundary between two adjacent foreign inclusions. The inclusion entity model is designed to restrict such spurious bracketing. It does so by way of another tree transformation. The new category FP (foreign phrase) is added below any node dominating at least one token marked FOM during training. For example, when encountering a FOM sequence dominated by PN as in Figure 5.5(a), the tree is modified so that it is the FP rule which generates the FOM tokens. Figure 5.5(b) shows the modified tree. In all cases, a unary rule PN→FP is introduced. As this extra rule decreases the probability of the entire tree, the parser has a bias to introduce as few of these rules as possible – thus limiting the number of categories which expand to FOMs. Once a candidate parse is created during testing, the inverse operation is applied, removing the FP node.

5.3.3 Method

For all experiments reported here, the different versions of the parser are trained on the TIGER treebank. As the inclusion and random sets are drawn from the whole treebank, it is necessary to ensure that the data used to train the parser does not overlap with these test sentences. The experiments are therefore designed as multi-fold cross-validation tests. Using 5 folds, each model is trained on 80% of the data while the remaining 20% is held out. The held-out set is then intersected with the inclusion set (or, respectively, the random set). The evaluation metrics are calculated on this sub-set of the inclusion set (or random set), using the parser trained on the corresponding training data. This process ensures that the test sentences are not contained in the training data.

The overall performance metrics of the parser are calculated on aggregated totals

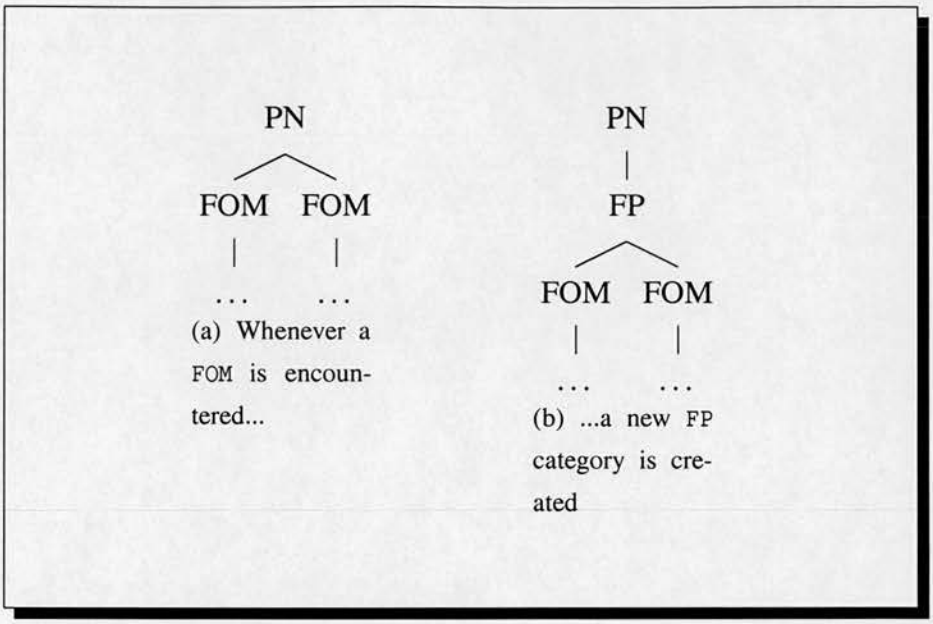


Figure 5.5: Tree transformation employed in the inclusion entity parser.

of the five held-out test sets. For each experiment, parsing performance is reported in terms of the standard PARSEVAL scores (Black *et al.*, 1991), including coverage (Cov), labelled precision (P) and recall (R) and F-score, the average number of crossing brackets (AvgCB), and the percentage of sentences parsed with zero and with two or fewer crossing brackets (0CB and $\leq$2CB). In addition, dependency accuracy (Dep) is also reported. Dependency accuracy is calculated by means of the approach described in Lin (1995), using the head-picking method employed by Dubey (2005a). The labelled bracketing figures (P, R and F) and the dependency score are calculated on all sentences, with those which are out-of-coverage getting zero counts. The crossing bracket scores are calculated only on those sentences which are successfully parsed.

Stratified shuffling is used to determine statistical difference between precision and recall values of different runs.[4] In particular, statistical difference is determined over the baseline and the perfect tagging model runs for both the inclusion and the random test sets. In order to differentiate between the different tests, Table 5.2 lists a set of diacritics used to indicate a given (in)significance.

[4]This approach to statistical testing is described in detail at: `http://www.cis.upenn.edu/~dbikel/software.html`

5.3.4 Results

5.3.4.1 Baseline and Perfect Tagging

The baseline, for which the unmodified parser is used, achieves a high coverage at over 99% for both the inclusion and the random sets (see Table 5.3). However, scores differ for the bracketing measures. Using stratified shuffling, a t-test on precision and recall affirms both to be significantly worse in the inclusion condition. Overall, the harmonic mean (F) of precision and recall is 65.2 on the random set, 6 points higher than 59.2 F observed on the inclusion set. Similarly, dependency and zero-cross-bracketing scores are higher on the random test set. The baseline parser produces on average 0.5 crossing brackets less per parsed random sentence than per inclusion sentence. These results strongly indicate that sentences containing English inclusions present difficulty for the parser, compared to length-matched sentences without inclusions.

When providing the parser with perfect tagging information, scores improve both for the inclusion and the random TIGER samples, resulting in F-scores of 62.2 and 67.3, respectively. However, the coverage for the inclusion set decreases to 92.7% whereas the coverage for the random set is 97.7%. In both cases, the lower coverage is caused by the parser being forced to use infrequent tag sequences, with the much lower coverage of the inclusion set likely due to infrequent tags (notable `FM`), associated with inclusions. While perfect tagging increases overall accuracy, the difference of 5.1 in F-score remains statistically significant between the random set and the inclusion set.[5] Although reduced over that of the equivalent baseline runs, this persisting difference shows that even with perfect tagging, parsing English inclusions is harder than parsing monolingual data.

So far, it was shown that the English inclusion classifier is able to detect sentences that are difficult to parse. It was also shown that perfect tagging helps to improve parsing performance but is insufficient when parsing sentences containing English inclusions. Next, it is examined how the knowledge provided by the English inclusion classifier can be exploited to improve parsing performance for such sentences.

[5]The average F-scores of all parsed sentences (ignoring failed parses) amount to 64.6 for the inclusion set and 68.1 for the random set to give an idea of how the coverage affects the F-scores in this experiments.

*	significantly different from inclusion baseline run
$\hat{*}$	not significantly different from inclusion baseline run
⋆	significantly different from perfect tagging inclusion run
$\hat{\star}$	not significantly different from perfect tagging inclusion run
○	significantly different from inclusion entity inclusion run
$\hat{\circ}$	not significantly different from inclusion entity inclusion run
#	significantly different from random baseline run
$\hat{\#}$	not significantly different from random baseline run
♯	significantly different from perfect tagging random run
$\hat{\sharp}$	not significantly different from perfect tagging random run
♮	significantly different from inclusion entity random run
$\hat{\natural}$	not significantly different from inclusion entity random run

Table 5.2: Meaning of diacritics indicating statistical (in)significance of t-tests using stratified shuffling compared to various runs.

Data	P	R	F	Dep	Cov	AvgCB	0CB	≤2CB
Baseline model								
Inclusion set	56.1#	62.6#	59.2	74.9	99.2	2.1	34.0	69.0
Random set	63.3*	67.3*	65.2	81.1	99.2	1.6	40.4	75.1
Perfect tagging model								
Inclusion set	61.3*♯	63.0$\hat{*}$♯	62.2	75.1	92.7	1.7	41.5	72.6
Random set	65.8#⋆	68.9#⋆	67.3	82.4	97.7	1.4	45.9	77.1
Word-by-word model								
Inclusion set	55.6$\hat{*}$⋆#♯	62.8$\hat{*}$⋆#♯	59.0	73.1	99.2	2.1	34.2	70.2
Random set	63.3*$\hat{\star}$$\hat{\#}$♯	67.3*⋆$\hat{\#}$$\hat{\sharp}$	65.2	81.1	99.2	1.6	40.4	75.1
Inclusion entity model								
Inclusion set	61.3*$\hat{\star}$$\hat{\#}$♯○	65.9*⋆$\hat{\#}$♯○	63.5	78.3	99.0	1.7	42.4	77.1
Random set	63.4*$\hat{\star}$$\hat{\#}$♯♮	67.5*⋆$\hat{\#}$$\hat{\sharp}$♮	65.4	80.8	99.2	1.6	40.1	75.7

Table 5.3: Baseline and perfect tagging results for inclusion and random sets and results for the word-by-word and the inclusion entity models.

5.3.4.2 Word-by-word Model

The word-by-word model achieves the same coverage on the inclusion set as the baseline but with a slightly lower F-score of 59.0. All other scores, including dependency accuracy and cross bracketing results are similar to those of the baseline (see Table 5.3). This shows that limiting the parser's choice of POS tags to those encountered for English inclusions is not sufficient to deal with such constructions correctly. In the error analysis presented in Section 5.3.5, the difficulty in parsing multi-word English inclusions in terms of recognising them as constituents, as opposed to recognising their individual POS tags, is examined in more detail. The aim is to overcome this problem with the inclusion entity model.

5.3.4.3 Inclusion Entity Model

The inclusion entity parser attains a coverage of 99.0% on the inclusion set, similiar to the coverage of 99.2% obtained by the baseline model on the same data. On all other measures, the inclusion entity model exceeds the performance of the baseline, with a precision of 61.3% (5.2% higher than the baseline), a recall of 65.9% (3.3% higher), an F-score of 63.5 (4.3 higher) and a dependency accuracy of 78.3% (3.4% higher). The differences in precision and recall between the inclusion entity model and the baseline model (both on the inclusion set) are statistically significant (t-test: $p \leq 0.001$ each). The average number of crossing brackets is 1.7 (0.4 lower), with 42.4% of the parsed sentences having no crossing brackets (8.4% higher), and 77.1% having two or fewer crossing brackets (8.1% higher). When testing the inclusion entity model on the random set, the performance is very similar to the baseline model on this data. While coverage is the same (99.2%), F and cross-bracketing scores are marginally improved, and the dependency score is marginally deteriorated. This shows that the inclusion entity model does not harm the parsing accuracy of sentences that do not actually contain foreign inclusions.

Not only does the inclusion entity parser perform above the baseline on every metric, its performance also exceeds that of the perfect tagging model on all measures except precision and average crossing brackets, where both models are tied. These results clearly indicate that the inclusion entity model is able to leverage the information about English inclusions provided by the English inclusion classifier. However, it is

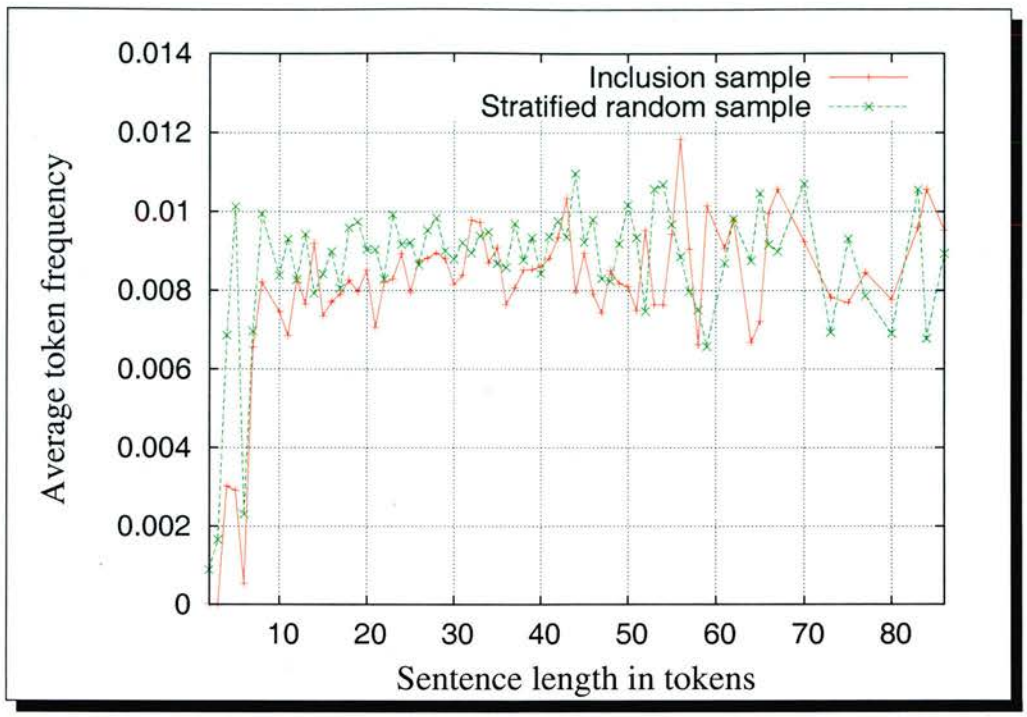


Figure 5.6: Average relative token frequencies for sentences of equal length.

also important to note that the performance of this model on the inclusion set is still consistently lower than that of the baseline model (2% in precision and 1.4% in recall, both not statistically significant: $\hat{\sharp}$), the perfect tagging model (4.5% in precision and 3% in recall, both statistically significant: $\sharp$) and the inclusion entity model (2.1% in precision and 1.6% in recall, both not statistically significant: $\hat{\natural}$) on the random data set. This demonstrates that sentences with inclusions are more difficult to parse with the treebank-induced parser than length-matched monolingual sentences, even in the presence of information about the inclusions that the parser can exploit.

Comparing the inclusion set to the length-matched random set is arguably not entirely fair as the latter set may not contain as many infrequent tokens as the inclusion set. Figure 5.6 shows the average relative token frequencies for sentences of equal length for both sets. It illustrates that the frequency profiles of the two data sets are broadly similar (the difference in means of both groups is only 0.000676, see Figure 5.4), albeit significantly different according to a paired t-test ($p \leq 0.05$). This difference is one reason that explains why the inclusion entity model's performance on the inclusion set does not reach the upper limit set by the stratified random sample.

Group	Mean	StDev
Inclusion set	0.008007	0.002140
Random set	0.008683	0.001781

Table 5.4: Means and standard deviations of frequency profiles for the inclusion and the stratified random set.

5.3.5 Error Analysis

The aim of this error analysis is to examine the exact reasons for why the inclusion entity model yields an improved performance over the baseline model when parsing sentences containing English inclusions. The error analysis is limited to 100 sentences selected from the inclusion set parsed with both the baseline and the inclusion entity model. This sample contains 109 English inclusions, five of which are false positives, i.e., the output of the English inclusion classifier is incorrect. The precision of the classifier in recognising multi-word English inclusions is therefore 95.4% for this TIGER sample. Before analysing the errors in the parsing output of the inclusion set in detail, it is worthwhile examining typical gold standard phrase categories of the multi-word English inclusions.

Phrase category	Frequency	Example
PN	91	The Independent
CH	10	Made in Germany
NP	4	Peace Enforcement
CNP	2	Botts and Company
–	2	Chief Executives

Table 5.5: Gold standard phrasal categories of English inclusions.

5.3.5.1 Gold Standard Phrase Categories

Table 5.5 lists the different types of phrase categories surrounding the 109 multi-word English inclusions in the error analysis sample and their frequency.[6] The last column lists a typical example for each category. The figures illustrate that the majority of multi-word English inclusions are contained in a proper noun (PN) phrase, including names of companies, political parties, organisations, films, books, newspapers, etc. The components of PN phrases tend to be marked with the grammatical function PNC, proper noun component (Brants *et al.*, 2002). A less frequent phrasal category of English inclusions is chunk (CH) which tends to be used for slogans, quotes or expressions like *Made in Germany*. The components of CH phrases are annotated with a grammatical function of type UC (unit component). Even in this small sample, phrase category annotations of English inclusions as either PN or CH, and not the other, can be misleading. For example, the organisation *Friends of the Earth* is annotated as PN, whereas another organisation *International Union for the Conservation of Nature* is marked as CH in the gold standard. The latter is believed to be an inconsistency in the annotation and should have been marked as PN as well.

The phrase category of an English inclusion with the syntactic function of a noun phrase which is neither a PN nor a CH is annotated as NP (noun phrase). One example is *Peace Enforcement* which is not translated into German and used rather like a buzzword in a sentence on UN missions. In this case, the POS tag of its individual tokens is NN (noun). The fact that this expression is not German is therefore lost in the gold standard annotation. Another example of an English inclusion NP in the gold standard is *Framingham Heart Study* which could arguably be of phrase category PN. Furthermore, the sample contains an example of phrase category CH, *Shopping Mall*, an English noun phrase. The least frequent type of phrase category used for English inclusions is CNP. In this sample, this category marks a company names made up of a conjunction, for example *Botts and Company*. The POS tags of the coordinated sisters are NE (named entity) and the English coordinated conjunction *and* is tagged as KON. Finally, there are also two cases, where the English inclusion itself is not contained in a phrase category. One of them is *Chief Executives* which is clearly an NP. These are believed to be annotation errors.

[6] All phrase category (node) labels and grammatical function (edge) labels occurring in the TIGER treebank annotation are listed and defined in Appendix C.

Phrase bracket (PB) frequency	BL	IE
$PB_{PRED} > PB_{GOLD}$	62%	51%
$PB_{PRED} < PB_{GOLD}$	11%	13%
$PB_{PRED} = PB_{GOLD}$	27%	36%

Table 5.6: Bracket frequency of the predicted baseline (BL) and the inclusion entity (IE) model output compared to the gold standard.

This analysis suggests that the annotation guidelines on foreign inclusions could be improved when differentiating between phrase categories containing foreign material. Despite the few inconsistencies and annotation errors discussed here, the large majority of English inclusions is consistently annotated as either PN or CH phrase. In the following, the errors in the parsing output of the inclusion set are examined in detail.

5.3.5.2 Phrase Bracketing

Table 5.6 summarises the number of phrase bracketing errors made for the inclusion set. For the majority of sentences (62%), the baseline model predicts more brackets than are present in the gold standard parse tree. This number decreases by 11% to 51% when parsing with the inclusion entity model. The baseline parser predicts fewer phrase brackets in the output compared to the gold standard in only 11% of sentences. This number slightly increases to 13% for the inclusion entity model. Generally, these numbers suggest that the baseline parser does not recognise English inclusions as constituents and instead parses their individual tokens as separate phrases. Provided with additional information of multi-word English inclusions in the training data, the parser is able to overcome this problem. This conclusion is further substantiated in the next section which examines parsing errors specifically caused by English inclusions.

5.3.5.3 Parsing Errors

In order to understand the parser's treatment of English inclusions, each parse tree is analysed as to how accurate the baseline and inclusion entity models are at predicting both phrase bracketing and phrase categories (see Table 5.7). For 46 inclusions (42.2%), the baseline parser makes an error with a negative effect on performance.

Errors	No. of inclusions (in %)	
Parser: baseline model, data: inclusion set		
Incorrect PB and PC	39	(35.8%)
Incorrect PC	5	(4.6%)
Incorrect PB	2	(1.8%)
Correct PB and PC	63	(57.8%)
Parser: inclusion entity model, data: inclusion set		
Incorrect PB and PC	6	(5.5%)
Incorrect PC	25	(22.9%)
Incorrect PB	4	(3.7%)
Correct PB and PC	74	(67.9%)

Table 5.7: Baseline and inclusion entity model errors for inclusions with respect to their phrase bracketing (PB) and phrase category (PC).

In 39 cases (35.8%), the phrase bracketing and phrase category are incorrect, and constituent boundaries occur within the inclusion, as is illustrated in Figure 5.7(a). When comparing this parsing output to the gold standard parse tree displayed in Figure 5.7(b), it is evident that the POS tagger did not recognise the English inclusion *Made In Heaven* as one proper name but rather as a named entity (NE) followed by a preposition (APPR), and followed by another NE. Most multi-word English inclusions contain tokens marked with the same POS tag in the gold standard, either all NE or FM. The POS tagger incorporated in the baseline parser fails to recognise this consistency within multi-word inclusions and often mistags at least one token as either common noun (NN), adjective (ADJA/ADJD), adverb (ADV), finite verb (VVFIN), irreflexive personal pronoun (PPER), preposition (APPR) or fused preposition and determiner (APPART). Such errors subsequently cause the parser to treat the constituents of inclusions as separate phrases. This leads to the prediction of constituent boundaries within inclusions and wrong phrase category and grammatical function assignment. It also has a detrimental effect on the parsing of the remainder of the sentence. Overall, the baseline model predicts the correct phrase bracketing and phrase category for 63 inclusions (57.8%).

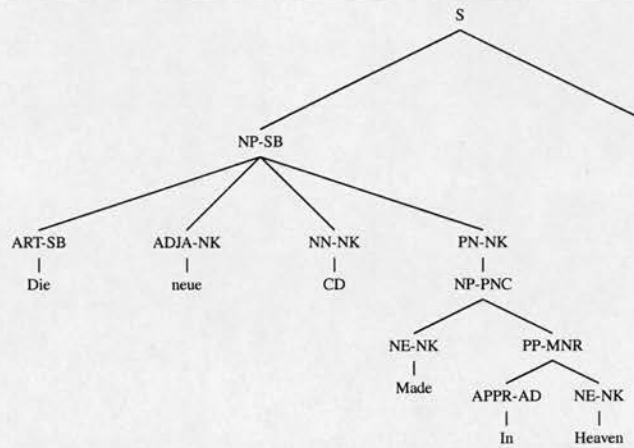
Conversely, the inclusion entity model, which is given information on tag consistency within inclusions via the FOM tags, is able to determine the correct phrase bracketing and phrase category for 67.9% of inclusions (10.1% more)[7], e.g. see Figure 5.7(c). Both the phrase bracketing and phrase category are predicted incorrectly in only 6 cases (5.5%). The inclusion entity model's improved phrase boundary prediction for 31 inclusions (28.4% more correct) is likely to have an overall positive effect on the parsing decisions made for the context which they appear in. Nevertheless, the inclusion entity parser still has difficulty determining the correct phrase category in 25 cases (22.9%) Unsurprisingly, the main confusion lies between assigning the categories PN, CH and NP, the most frequent phrase categories of multi-word English inclusions. As explained in Section 5.3.5.1, this is also partially due to the ambiguity between these phrases in the gold standard. For example, the English organisation *International Union for the Conversation of Nature* was predicted to be a proper noun phrase by the inclusion entity parser, as one would expect. However, as this organisation is marked as a chunk phrase in the gold standard, the parser's phrase category prediction has a negative effect on the F-score in this case. However, as the phrase bracketing is the same as in the gold standard, such errors do not affect bracketing scores negatively.

Finally, few parsing errors (4) are caused by the inclusion entity parser due to the markup of false positive inclusions, mainly as a result of boundary errors. For example, the English inclusion classifier failed to recognise the word *Fast* as part of the English inclusion *Fast Times at Ridgemont High*, which caused the parser to make the mistake shown in Figure 5.8.

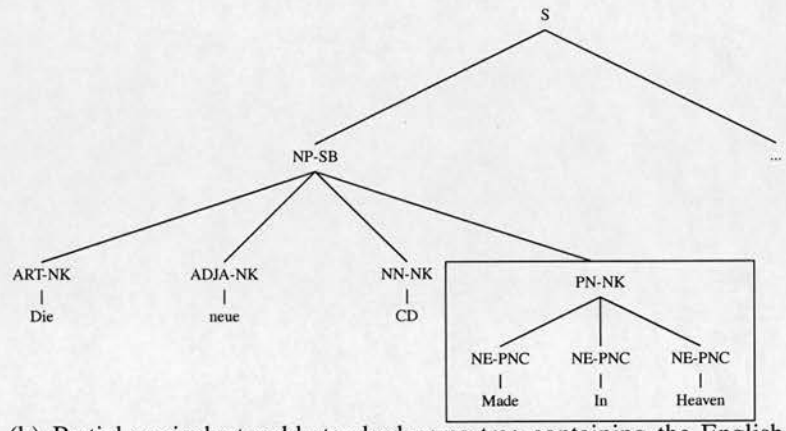
5.3.6 Discussion

As English inclusions occurring in German text are the cause for increasing language mixing, this chapter started with the hypothesis that such inclusions can be a significant source of error for monolingual parsers. Evidence for this hypothesis was provided by showing that a baseline model, i.e. an unmodified treebank-induced parser for German, performs substantially worse on a set of sentences with inclusions compared to a set of length-matched sentences randomly sampled from the same corpus. A perfect tagging

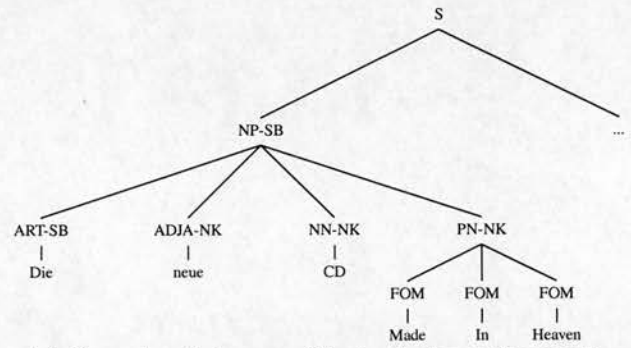
[7]Differences refer to percentage points.



(a) Partial parse of the baseline model containing a constituent boundary in the English inclusion. Translation: *The new CD Made In Heaven ...*



(b) Partial equivalent gold standard parse tree containing the English inclusion.



(c) Correct partial parse of the inclusion entity model.

Figure 5.7: Partial parse trees of produced by the baseline parser, found in the gold standard and output by the inclusion entity model.

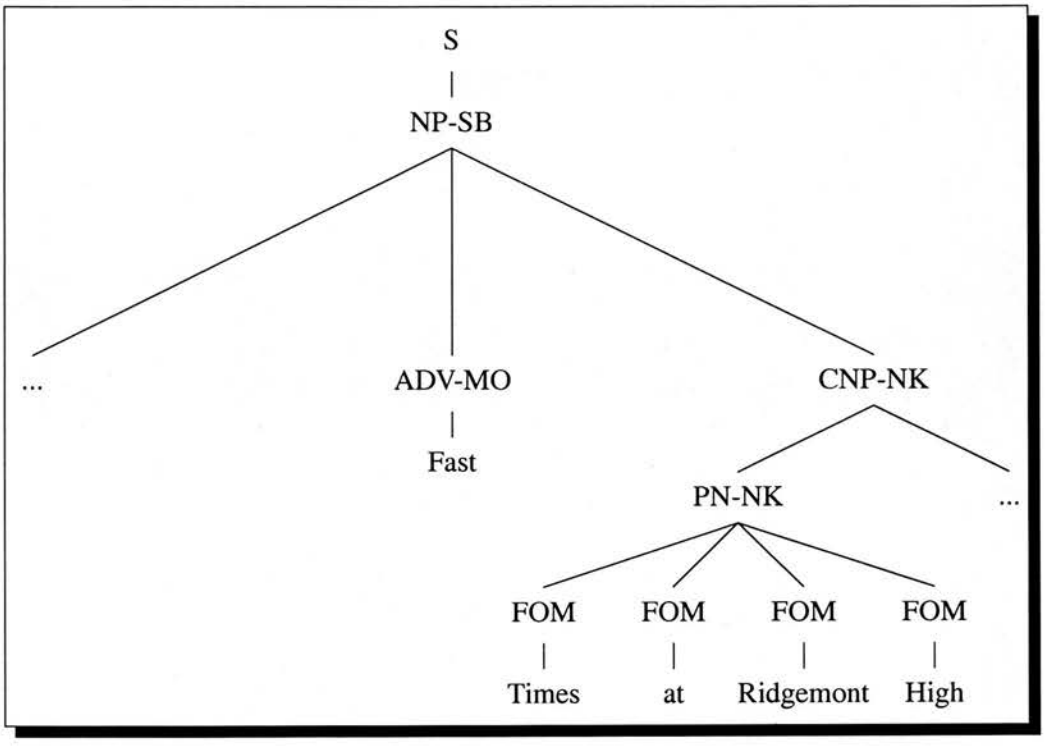


Figure 5.8: Partial parse of the inclusion entity model for a false positive inclusion.

model was also investigated in which the parser is given gold standard POS tags in the input. Even under these conditions, parsing performance is substantially lower on the inclusion set than on the randomly sampled set. The English inclusion classifier is essentially able to spot sentences that are difficult to parse.

To address the problem of inclusions in parsing, the English inclusion classifier, an annotation-free method for accurately detecting inclusions which compares favourably against a supervised ML approach, was run over the German TIGER treebank. Two methods for interfacing inclusion detection with parsing were tested. The first method restricts the POS tags of inclusions that the parser can assign to those found in the data. The second method applies tree transformations to ensure that inclusions are treated as phrases. An evaluation on the TIGER corpus shows that the second approach achieves a performance gain of 4.3 points F-score over a baseline of no inclusion detection, and even outperforms a model with access to perfect POS tagging of inclusions.

To summarise, it was shown that foreign inclusions present a problem for a monolingual treebank-induced parser. It appears that it is insufficient to know where inclusions are or what their parts of speech are. Parsing accuracy only improves if the parser also has knowledge about the structure of the inclusions. It is particularly important to know when adjacent foreign words are likely to be part of the same phrase. The error analysis showed that this prevents cascading errors further up in the parse tree.

The results indicate that future work could improve parsing performance for inclusions further as parsing the inclusion set is still harder than parsing a randomly sampled set, even for the best-performing model. This marks an upper bound on the performance expected from a parser that uses inclusion detection. The next section will evaluate the English inclusion classifier's merit when applied to parsing with a hand-crafted grammar.

5.4 Parsing Experiments with a Hand-crafted Grammar

A second set of parsing experiments involve a German parser based on a hand-crafted grammar, using the Lexical Functional Grammar (LFG) formalism, developed at the University of Stuttgart. The nature of parsing German sentences containing English inclusions with this monolingual parser will be analysed in detail. The aim is to determine if inclusions pose as much difficulty as they do with a monolingual treebank-induced parser and to test if additional knowledge about this language-mixing phenomenon can be exploited to overcome this problem. Considering that the treebank-induced parser sees at least some inclusions in the training data, although they are sparse, a hand-written symbolic parser is expected to have even more difficulty in dealing with English inclusions as it generally does not contain rules that handle foreign material. Before focussing on the experiments, the parser is briefly introduced.

5.4.1 Parser

The Xerox Linguistic Environment (XLE) is the underlying parsing platform used in the following set of experiments (John T. Maxwell and Kaplan, 1993). This platform functions in conjunction with a hand-written large-scale LFG of German developed by Butt *et al.* (2002) and improved, for example, by Dipper (2003), Rohrer and Forst (2006) and Forst and Kaplan (2006). The version of the German grammar used here contains 274 LFG style rules compiled into an automaton with 6,584 states and 22,241 arcs. Before parsing, the input is firstly tokenised and normalised. Subsequently, string-based multi-word identification is carried out, followed by morphological analysis, analysis guessing for unknown words and lexically-based multi-word identification (Rohrer and Forst, 2006; Forst and Kaplan, 2006). Forst and Kaplan (2006) improved the parsing coverage for this grammar from 68.3% to 73.4% on sentences 8,001 to 10,000 of the TIGER corpus by revising the integrated tokeniser.

The parser outputs Prolog-encoded constituent-structure (c-structure) and functional-structure (f-structure) analyses for each sentence. These two representation levels are fundamental to the linguistic theory of LFG and encode the syntactic properties of sentences. For in-depth introductions to LFG, see Falk (2001), Bresnan (2001), Dalrymple (2001) and Dalrymple *et al.* (1995). While c-structures represent the word order and phrasal grouping of a sentence in a tree, f-structures encode the

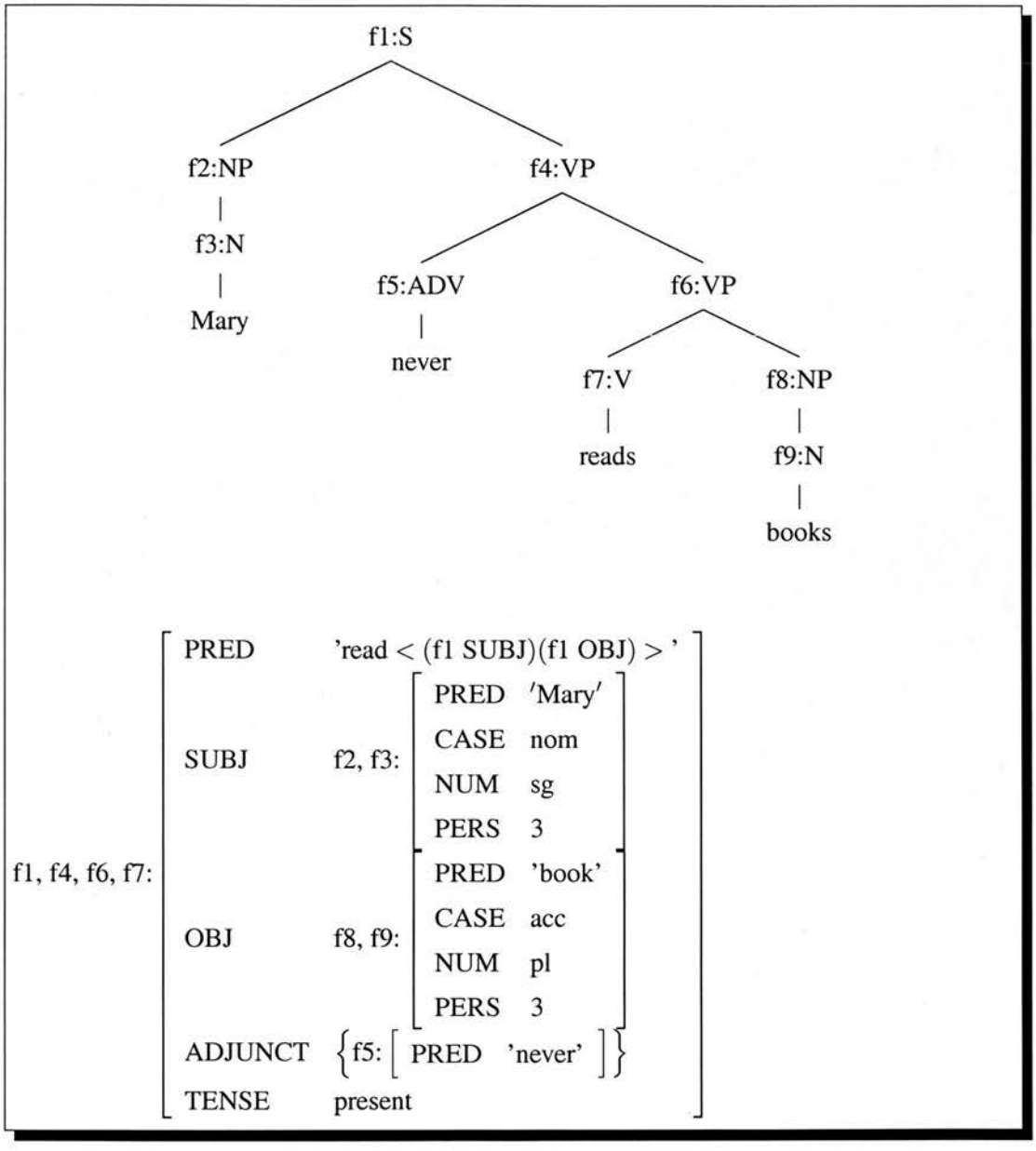


Figure 5.9: Complete c- and f-structures for an English example sentence (Dipper, 2003).

grammatical functions of sentence constituents. The complete c- and f-structures of the sentence *Mary never reads books.* are illustrated in Figure 5.9. The mapping relation from the c- to the f-structure is obtained by means of f-structure variables (f1-9).

5.4.2 Parsing Modifications

The data sets used for the experiments with the LFG parser are the same as those parsed with the treebank-induced parser (see Section 5.2.1): the inclusion set and the stratified random set, both samples from the TIGER corpus. Two variations of the inclusion set are presented to the LFG parser: (1) a raw text version, and (2) a version marked for multi-word inclusions. As previously, the LFG parser does not process foreign inclusions in any special way and is therefore most likely to treat them in the same way as rare or unseen words. Parsing with marked-up multi-word inclusions is inspired by the hypothesis that the parser fails to parse inclusions correctly as it is unable to recognise adjacent inclusions as constituents of the same phrase.

5.4.2.1 Multi-word Inclusion Parsing

For the baseline, the parser is simply presented with the raw sentences themselves (see Example (1)). Multi-word inclusion parsing, however, involves an adaptation which allows for strings to be surrounded by the element <mwe> as multi-word expressions. This additional markup is expected to be equally useful for parsing English inclusions. The inclusions in the inclusion set (as detected automatically by the classifier) are therefore marked with this additional element in the input for the multi-word inclusion parsing (see Example (2)). As the English inclusion classifier does not perform with perfect accuracy, this test suite contains some erroneous mark-up. However, the error analysis on a sub-set of the treebank-induced parsing output showed that the classifier performs at a high precision of 95.4% (see Section 5.3.5). Whenever a sequence of tokens is marked with the <mwe> element, the parser treats this sequence as a phrase.

(1) **Raw:** Dann wird auch in der Economy Class auf Langstreckenflügen eine Menükarte gereicht .

(2) **Marked:** Dann wird auch in der <mwe>Economy Class</mwe> auf Langstreckenflügen eine Menükarte gereicht .

Translation: *Then a menu is also handed out in the economy class during long-haul flights.*

5.4.3 Method

Apart from the additional inclusion mark-up, the baseline and the multi-word inclusion parsing runs are both carried out under exactly the same conditions. As recommended by the developers of the parser, the variables `timeout` and `max_xle_scratch_storage`, which specify the usage of time (in CPU seconds) and the storage space (in megabytes) after which the parsing action of a sentence is aborted, are set to 200 and 1700, respectively.

All LFG parsing experiments are currently only evaluated in terms of their coverage. Labelled precision and recall etc. are not currently calculated as the work on designing transfer rules that convert f-structures into dependency triples is still ongoing. Once this work is completed, the output of the LFG parser can be compared against the corresponding gold standard TIGER dependency triples, using the triples matching mechanism that is part of the XLE platform. This will make it possible to determine labelled precision and recall on dependency triples. As the LFG parser often outputs several solutions which are ranked according to likelihood only in a post-processing stage, the evaluation on dependency triples can be done in three different ways. The first two options are evaluating the solution which is closest to the gold standard and a randomly selected parse of multiple solutions which provide an upper bound and a baseline performance of the parser, respectively. The third option is using a stochastic disambiguation component that determines the most likely solution which is then used for evaluation. The latter approach is expected to return scores in-between those attained for the best and the randomly selected derivation. All three types of evaluation can be done directly on the Prolog output which will save time re-running the experiments.

The first set of experiments compares the coverage of the LFG parser for the inclusion set to that of the same parser for the stratified random set. In the second set of experiments, baseline and multi-word inclusion parsing are compared on the inclusion set, i.e. once in raw and once in marked-up format. All results are shown in Table 5.8.

Parsing	Baseline				Multi-word inclusion	
Data Set	Random set		Raw inclusion set		Marked inclusion set	
Successful parse	321	53.9%	208	34.9%	457	76.7%
One derivation	37	11.5%	21	10.1%	51	11.2%
Several derivations	284	88.5%	187	89.9%	406	88.8%
No parse	275	46.1%	388	65.1%	139	23.3%
Parser failed	140	50.9%	260	67.0%	2	1.4%
Time/storage out	135	49.1%	128	33.0%	137	98.6%
Total	596	100%	596	100%	596	100%

Table 5.8: LFG parsing results for baseline and multi-word inclusion parsing.

5.4.4 Results

5.4.4.1 Baseline Parsing

The scores listed in Table 5.8 show that baseline parsing performs with a low coverage of 34.9% on sentences containing English inclusions. This data set has a high average sentence length of 28.4 tokens. On the stratified random test suite with the same average sentence length, parsing coverage is higher at 53.9%. In order to get an idea about the LFG parser's general performance, a coverage of 72.7% is reached on a completely random TIGER set with an average sentence length of 17.6 tokens (not displayed in Table 5.8).

These results demonstrate that the baseline parser struggles to perform well in terms of parsing coverage on data containing English inclusions. In fact, its coverage more than doubles when evaluated on completely random data. Even when controlling for sentence length, the parser achieves 19% higher coverage on the stratified random set. The hypothesis that sentences containing multi-word English inclusions pose difficulty to the parser is therefore correct. The next experiment investigates if the additional language knowledge generated by the English inclusion classifier can improve upon the low coverage obtained by the baseline parser.

5.4.4.2 Multi-word Inclusion Parsing

Given the raw inclusion set, the baseline parser produces at least one successful parse for 208 out of 596 sentences (34.9%). In 10.1% cases it yields one derivation and in 89.9% cases it predicts multiple derivations. On average, the baseline parser produces 275 analyses per sentence for the inclusion set. It is unable to produce a derivation for 65.1% of sentences from the inclusion set either because it fails or due to a time or storage out. Conversely, additional multi-word inclusion markup results in 76.7% of successful parses for the marked inclusion set (41.8% more than the baseline parser), 11.2% of them (1.1% more) with one derivation and 88.8% with several derivations. On average, this parser produces 406 analyses per sentence. However, the average number of derivations for sentences which are also parsed successfully by the baseline parser is only 154, 44.2% less than produced by the baseline parser itself (276). Multi-word inclusion parsing not only outperforms the coverage of the baseline parser, and for that matter the coverage obtained for the length-matched random set (53.9%) and the completely random set (72.7%), but also produces less ambiguous parsing output. Furthermore, it only fails to parse one sentence for which the baseline parser manages to produce derivations.

5.4.5 Discussion

The results show that the hand-crafted LFG parser also struggles to deal with German sentences containing multi-word English expressions. By providing the parser with the additional knowledge of where in the data English multi-word expressions are, coverage increases by 41.8% on sentences containing multi-word inclusions. The hypothesis that the additional language knowledge for multi-word inclusions provided by the classifier improves parsing coverage can therefore be regarded as correct. Given that the average number of derivations obtained per sentence decreases, the additional mark-up allows the multi-word inclusion parser to disambiguate between several derivations already predicted by the baseline parser. At this point, it is unclear how accurately the parser performs for all successful parses and how effective a stochastic disambiguation between multiple solutions would be. This additional evaluation remains to carried out as future work.

The multi-word inclusion markup helps both parsers partially with the identification of proper noun phrases, and partially with the detection of other multi-word expressions. As a result, it would be interesting to derive a way of interfacing the parser either with a named entity recogniser or a multi-word expression extractor, in addition to the English inclusion classifier, in order to determine the difference in parsing performance on the stratified random and inclusion sets.

5.5 Chapter Summary

This chapter started out with the hypothesis that English inclusions can constitute a significant source of error for monolingual parsers. Evidence for this hypothesis was provided both for a treebank-induced parser and a rule-based grammar parser for German. It was shown that the English inclusion classifier is able to detect sentences that are difficult to parse, or which the parser is unable to parse. When interfacing English inclusion detection with parsing, ensuring that inclusions are treated as phrases, parsing performance increases significantly for the treebank-induced parser. When conducting a similar experiment with the rule-based grammar parser, coverage increases substantially and ambiguity decreases. This shows that foreign inclusions present a non-negligible problem to monolingual parsers. Their correct detection avoids the cascading of errors in the remainder of the parse tree. As parsers are integrated into many NLP applications, including text-to-speech synthesis or machine translation, improved parsing of foreign material in mixed-lingual text will also have positive effects on the performance of any subsequent language processing. The direct benefit of English inclusion detection to such applications is investigated in the next chapter.

Chapter 6

Other Potential Applications

This chapter discusses in detail three other applications or fields, namely **text-to-speech synthesis**, **machine translation** as well as **linguistics and lexicography**, for which the automatic identification of foreign, in particular English, inclusions would be beneficial. In Section 6.1, the benefit and implications of applying English inclusion detection as a pre-processing step in a text-to-speech (TTS) synthesiser are discussed in greater detail and a strategy for an extrinsic evaluation of the inclusion classifier is proposed. This section includes detailed reviews of lessons learnt from production and perception experiments with mixed-lingual speech and conclusions from studies in the field of second language acquisition (SLA). Reviews of research efforts on all aspects of synthesising speech containing foreign inclusions are also presented in detail. Sections 6.2 and 6.3, which summarise the value of English inclusion detection for machine translation as well as linguistics and lexicography, are less detailed but present several ideas to pursue in future work.

6.1 Text-to-Speech Synthesis

In today's globalised world, real-world TTS applications must handle a variety of texts in languages which sometimes cannot be predicted in advance. These can include technical documents or international communication mainly in one language and partially in a second, as well as messages containing foreign names or expressions. This section first presents production and perception studies undertaken to investigate the pronunciation of foreign inclusions in various languages as well as in the related research field of SLA. Following an overview of TTS synthesis and its evaluation, this section describes existing studies tackling aspects of polyglot speech synthesis. Their paradigms are to some extent modelled on the theories of the production and perception of foreign speech sounds. Based on this review, this section finally envisages ways to evaluate the merit of using English inclusion detection in the pre-processing of a TTS synthesiser to investigate how the quality of the output is affected given the additional language knowledge. Conducting this task-based evaluation is left to future work.

6.1.1 Pronunciation of Foreign Words

Before exploring efforts in the synthesis of mixed-lingual text, it is useful to gain an insight into the production and perception of foreign speech sounds by speakers of a particular language. A speaker may be inclined to pronounce foreign words embedded in text written predominantly in their own language differently than when the same word appears surrounded by text in the foreign language. The degree of adaptation of the pronunciation of foreign inclusions to the phonological system of the speaker's native tongue is dependent on a series of factors, amongst others co-articulation, economy of effort, age, fluency in the foreign language, the capability of rendering the pronunciation as well as the frequency of a particular foreign inclusion in society. Moreover, psychological and social factors can play a role in choosing a particular pronunciation. For example, adopting the correct foreign pronunciation may give the impression of an exaggerated level of sophistication and consequently be disapproved of in common language use. On the other hand, a speaker may expect a certain competence from his listeners. While some factors are concrete and lend themselves well to production and perception studies, others are more intangible and difficult to model or control for. Several studies have been conducted which shed light on some of these aspects.

6.1.1.1 Foreign Speech Sounds in Polyglot Speech

This section summarises the results of several production and perception studies with German, English and Swedish subjects investigating the pronunciation of anglicisms and English proper names. Their findings not only provide a better understanding of the pronunciation of foreign words but also identify various factors that affect it.

Viereck (1980) analysed the use and comprehension of 42 anglicisms by German speakers and found that older subjects and those without English language skills adjust the pronunciation of anglicisms to German more than younger ones and those who were able to speak English. Fink (1980) and Glahn (2000) report two further production studies involving the pronunciation of anglicisms and English names as produced by German subjects in order to determine the percentage of words pronounced as English or non-English. Fink's study is based on anglicisms and English brand names that were not uncommon in German at the time of the experiment and which could therefore be pronounced according to English pronunciation patterns. 51 such stimuli were presented in a list of words without context to 184 subjects of different professions, ages and gender who were asked to read them. Analysing the pronunciation of each word, Fink then classified it on the word level as being either English, German or mixed. His results show that 63% of words are pronounced according to English, 25% with German and 11% with mixed pronunciation patterns. However, no consideration is given to the pronunciation of individual phones. Fink also reports that the English language knowledge of subjects influences their pronunciation of English words whereas gender does not. In contrast, Glahn (2000) who analysed the use of English loans in broadcasts of two German public TV channels finds that the majority (64.4%) are rendered with non-English pronunciations. A pronunciation was considered non-English if at least one of its sounds was non-English. In addition, Glahn provides a list of English sounds contained in his data along with their pronunciations in the recordings. While proper names were excluded from his analysis, Glahn considered any type of loan, including semantic and syntactic influences of English on German which make both studies difficult to compare. Considering that English words can be produced with mixed-lingual pronunciation patterns, labelling language pronunciation on the word level is not very conclusive or useful particularly when the aim is to synthesise text.

Several research efforts have therefore focused on the rendering of specific English sounds in German and other languages (e.g. Jabłoński, 1990; Busse, 1994; Greisbach, 2003). Jabłoński, for example, shows that the most common phenomena of adapting English sounds to native pronunciations are phone substitutions, elisions, epenthesis or nativised stress patterns. Busse (1994) examined the pronunciation of anglicisms containing the letter *u* as well as those beginning with the phones [ʤ] or [st], or containing [ɒ]. Besides age and English language skills of subjects, he found that the pronunciation of anglicisms is also highly influenced by their origin (i.e. British or American English), their orthographic integration into the receiver language as well as their popularity. He established that older anglicisms are almost completely adapted to German whereas newer ones and those with a very specific meaning are pronounced with English phones. Greisbach (2003) examined how German speakers pronounce the voiced palato-alveolar affricate [ʤ] in English words and French nasal vowels in words of French origin. According to his results, the pronunciation of the English affricative depends on the speaker's age but not their educational background.

Some studies limit their analysis to proper names (e.g Fitt, 1998). Fitt conducted extensive production and perception experiments to investigate the adaptation of foreign names in English. Subjects were asked to produce stimuli of six language origins in spoken or written form when presented to them as text or in recordings. Even though many subjects were able to guess the language origin of different stimuli, they only produced the non-English sounds [œ] and [ʁ] from the German phone inventory correctly. Abresch (2007) concludes from this finding that native English speakers accept or are aware of foreign phone segments to a much lesser degree than German native speakers with respect to English phones.

Eklund and Lindström (1996, 1998, 2001), Lindström and Eklund (1999a,b, 2000, 2002) and Lindström and Kasaty (2000) have studied the production of foreign speech sounds in Swedish and its multi-dimensional implications for speech recognition and synthesis. They hypothesise that some foreign speech sounds are commonly used in every-day Swedish by the majority of the population. They recorded a set of sentences containing English speech sounds which are normally not included in the Swedish phonological system and do not have phonemic or allophonic functions. Eklund and Lindström (1998) introduce the term **xenophones** for such phones. The sentences were read out by 460 subjects aged between 15 and 75, living all over Sweden.

While all subjects were able to produced the foreign target vowels and diphthongs [aɪ], [eɪ], [əʊ], [juː] and [æ] successfully in 90% of cases or more, their approximation varied for different consonants. The speakers adjusted most foreign phone consonants to a similar one in their native tongue. Practically no subject succeeded in producing the voiced alveolar fricative [z] in *mu̲sic* or the post-alveolar fricative [ʒ] in *televi̲sion* and opted instead for almost full adjustment to Swedish speech sounds. The same behaviour was observed for the English [ɬ], e.g. in *worl̲d*, which is actually quite dissimilar sounding to the Swedish [l]. In the same way, the majority of subjects adjusted the English [w] in *we* to a Swedish [v]. Conversely, the voiceless affricative [tʃ] in *Baywatch* was produced correctly by the large majority of speakers. Moreover, the voiced [ð] and unvoiced [θ] dental fricatives appearing in *the* and *thriller* respectively were produced by close to 50% of all subjects even though no similar speech sounds exist in Swedish.

These results indicate that, when reading English words embedded within Swedish sentences, many Swedish speakers adjust some English xenophones to Swedish but added others to their phone inventory which they are able to approximate successfully to varying degrees. Similarly, Trancoso *et al.* (1999) conclude from experiments in which German speakers were asked to pronounce French place names, and vice versa, that many, even inexperienced speakers of the foreign language are able to produce sounds outside their native language phone inventory. They either attempt to approximate the pronunciation of the respective foreign language or another related foreign language which they are more experienced in (in this case English).

Lindström and Eklund (1999a) also analyse the results of their production study in terms of age, gender and regional dialect variations. They show that gender and regional differences between groups do not significantly influence the production of foreign speech sounds. However, age does play a role as expected. The majority of subjects in the age groups 16-25 and 26-35 produce foreign phones correctly in most cases. The youngest subjects, aged 15 and younger, often fully adjust foreign phones to Swedish ones. This behaviour is attributed to the lower cultural exposure due to their young age. This research is based on production rather than perception or evaluation. Their reason for opting for this approach is that it shows people's attitudes towards the occurrence of foreign inclusions in a more subconscious manner than, for example, in an evaluation of the quality of different versions of synthesised speech.

Eklund and Lindström (2001) conclude that there is a clear correlation between increasing educational level and closer approximation of the foreign language pronunciation. A high educational standard of subjects correlates with a high degree of xenophone inclusions. On the other hand, Ljung (1998)[1] reports negative attitudes towards foreign expressions correlating with a high level of education. Lindström and Eklund (2002) argue that this discrepancy can be explained by the fact that Ljung (1998) bases his observations on subjects who understood the purpose of the experiment. When being asked explicitly to read a sentence containing foreign inclusions, highly educated people may make a conscious decision to preserve their native tongue from foreign influence and adapt the inclusion's pronunciation to patterns of their own language. The same highly educated subjects might be more prone to render foreign inclusions according to the source language when unaware of the scenario.

Lindström and Eklund (2000) also infer from their findings that a Swedish TTS system must be capable of producing the appropriate pronunciation of foreign names and words in running Swedish texts, as users would have difficulties in accepting a TTS system with a lower level of competence than their own. Does this mean that Swedish listeners would prefer a TTS system which at least produces those foreign speech sounds correctly that are not included in the phonological inventory of their native language? How would they react to a system that pronounces all foreign speech sounds, including those with similar counterparts in the speaker's language, authentically according to the phonetic and acoustic characteristics of the foreign language? Would highly educated people more easily accept such synthesised speech? How would listeners perceive a system that adjusts all foreign speech sounds to sounds in their native language? Provided that this type of synthesis output is intelligible, would less highly educated listeners find it easier to accept it? Only a controlled and detailed perception study in terms of naturalness, intelligibility, pleasantness etc. would assist in evaluating the benefit of adding certain xenophones into the inventory of a TTS system. Lindström and Eklund (2000) do hypothesise that a low inclusion level of xenophones may not appear primarily in the intelligibility dimension, but portray itself to the listener as a synthesiser with a low level of education. On the other hand, if too many xenophones are added, some users might be at a disadvantage, particularly with regards to non-English foreign inclusions.

[1] Quoted in Lindström and Eklund (2002)

The pronunciation of English inclusions occurring in another language has also been studied by Abresch (2007). She conducted two extensive production experiments to test the extent to which English xenophones are nativised by German native speakers in German and English contexts as well as a perception study to test the conclusions drawn from the production study. 40 subjects (22 female and 18 male) aged 16 to 82 and with various educational backgrounds and English language skills participated in the first production experiment. They were asked to read German sentences with embedded anglicisms or English proper names containing English xenophones (including consonants, vowels and diphthongs). The carrier sentences containing the anglicisms and English names were selected from the online corpus *Deutscher Wortschatz*.[2] The majority of anglicisms were first documented in the latter half of the last century. One selection criterion was that they should not be considerably integrated into German so as to allow a potential English pronunciation. Abresch also included pseudo-anglicisms in her sample.

The results of this study show that there is a great variability in the realisations of English xenophones by German speakers. While 62.4% of xenophones were substituted by German phones, 37.6% of them were articulated like the original English phone either with a British or an American rendering. English diphtongs and consonants ([əʊ], or [oʊ], [ɛɪ], [θ], [ð], [ʤ] in onset position, [ɹ], [w], [s], [sp] and [st]) tended to be pronounced correctly with the exception of voiced obstruents ([b], [d], [g], [dz], [ʤ], [z] and [v] all in coda position) and the velarised [ɬ]. They and English vowels ([æ], [ɑː], [ʌ], [ɛə], [ɪə], [ʊə], [ɜː] and [ɒ]) tended to be substituted by German phones. The extent of this substitution process is highly influenced by the age and English language skills of the subjects. Older subjects and those who are less skilled in English substituted on average more English phones. Moreover, xenophones in proper names tended to be less often substituted and more often rendered like the original. However, gender was not found to have an effect on the pronunciation of English xenophones. These findings are similar to those made by Eklund and Lindström (2001), when reading English words embedded within German sentences, many German speakers substitute some English xenophones by German phones but add others to their phone inventory. While age and educational standard are influencing factors in this process, gender does not play a role.

[2] `http://wortschatz.uni-leipzig.de`

The second production study involved the pronunciation of anglicisms in German and English carrier sentences in order to test whether language context affects the substitution rate of English xenophones by subjects with excellent English language skills. While they substituted 61.9% of English xenophones in anglicisms embedded in German carrier sentences, they only substituted 31.5% when the same anglicisms appeared in English sentences. Therefore, anglicisms are nativised in German context even when the speakers' knowledge of English is advanced which means that some degree of nativisation is independent of the foreign language skills of the speaker.

Abresch (2007) concludes with a perception study to test whether the information learnt in the first production study corresponds to listeners' preference. 50 subjects aged between 16 and 75 and with different English language skills participated in this test. They were asked to listen to German sentences with embedded anglicisms and English names that contained the same English xenophones investigated in the first production experiment or corresponding substitutions produced by the subjects. Subjects then had to rate the variations of each sentence according to preference. Abresch determined that some English xenophones are clearly preferred over their German substitutions. With few exceptions, this group largely coincides with the xenophones that were also more often rendered closely to their original in the production study. Even subjects with no or little knowledge of English preferred certain xenophones over German substitutions. No significant difference was found in the preference of xenophones over substitutions in anglicisms and proper names. Abresch also shows that British renderings of English xenophones are mostly preferred over American ones.

While it is not possible to draw cross-linguistic conclusions of how anglicisms and English proper names are pronounced in other languages per se, the latter studies clearly show that their pronunciation patterns vary from those of both the receiver and donor languages. Further insight into this issue can be gained by examining studies in the area of SLA.

6.1.1.2 Foreign Speech Sounds in Second Language Acquisition

The problem of how speakers of a native language (L1) perceive and produce sounds in a foreign language (L2) is central to the research of SLA. This field differentiates between identical, similar and new L2 sounds when drawing conclusions with regard to their pronunciation in L1. An **identical L2 sound** is represented by the same IPA

symbol used to represent a sound in L1 and there is no significant acoustic difference between the sound pair. A **similar L2 sound** is represented by the same IPA symbol as an L1 sound, even though statistical analyses reveal significant acoustic differences between the two. A **new L2 sound** differs acoustically and perceptually from the sounds in L1 that most closely resemble it. Unlike a similar sound, a new sound is represented by an IPA symbol that is not used for L1 sounds (Flege, 1997).

For example, Flege (1987) argues that language learners place an L2 sound that is identical or sufficiently similar to an L1 sound in the same phonetic category, a cognitive mechanism called **equivalence class**. This process prevents adult learners from forming a new category for such L2 sounds and results in an accented pronunciation even after a lengthy exposure to L2. Evidence for this theory was found by Bohn and Flege (1992) for the English vowels [i] and [ɪ] (e.g. in the words *beat* and *bit*) when pronounced by German native speakers with varying degrees of English language experience. Bohn and Flege conclude that this behaviour can be interpreted according to the general principle of least effort. Although experienced and inexperienced speakers largely retain properties of the German [i] and [ɪ] in their English pronunciations, Flege's listening experiment shows that they are as highly intelligible to English native listeners as the English [i] and [ɪ] produced by English native speakers. Flege therefore suggests that phonetic learning of similar sounds takes place during early L2 exposure and does not improve considerably when gaining more L2 experience.

Flege and colleagues have also carried out extensive research on the production and perception of L2 sounds that lack a similar sound in L1, in other words new sounds or xenophones. According to their theory, adult learners will produce new L2 sounds more authentically (than similar L2 sounds) with extended L2 exposure as they establish new phonetic categories for such sounds. Bohn and Flege (1992) demonstrated that experienced but not inexperienced German learners of English produce the English [æ] (e.g. in *bat*) close enough to the English acoustic form, a sound that has no counterpart in German. Flege further indicates that a similar L2 sound which is in close proximity to new a L2 sound (for which the inexperienced foreign language learner has not yet established a phonetic category) in the acoustic vowel space will be produced closer to the acoustic norm of L2. However, once learners are able to produce an acoustic contrast between the similar and the neighbouring new sound, they tend to produce the similar sound with the acoustic characteristics of L1. Flege has identified

this effect both for German speakers where the similar English sound is [e] and the new English sound is [æ] (Bohn and Flege, 1992) as well as for French speakers, the similar English sound being [u] and the new English sound [y] (Flege, 1987). In Flege *et al.* (1997), the work on German learners of English (Bohn and Flege, 1990, 1992) is extended to native speakers of Spanish, Mandarin and Korean, three typologically diverse languages. Although the conclusions are similar, the paper stresses that further research is required to determine the degree of perceived relatedness between vowels found in English and in each of the native languages examined in the experiment.

Although similar in some respects, SLA research does not deal with the exact same issues involved in the production and perception of embedded foreign inclusions. It addresses the way in which speakers of a language L1 approach a language L2 with the intention of mastering it fluently, while sometimes living in a country where L2 is spoken. The research presented in this thesis, on the other hand, focuses on the occurrence of lexical items from a foreign language embedded within utterances in a native language and on how such foreign inclusions are dealt with by native speakers in their own language surrounding. In fact, the finding of SLA studies that similar L2 sounds are pronounced with properties of L1 and are therefore adapted is likely to be even more severe for embedded foreign inclusions as a result of co-articulation and a series of other factors determined in Section 6.1.1.1.

6.1.1.3 Implications for the Perception of Polyglot TTS

In the context of polyglot TTS synthesis, first and foremost the type and degree of adjusting foreign inclusions to the native language must be addressed. As the studies described in Section 6.1.1.1 have illustrated, pronunciation of foreign inclusions by native speakers varies depending on a series of factors, including age and educational level of the speaker. Other aspects that play a role in the production of foreign inclusions involve the orthography, the context of the particular foreign inclusion, the effort of production as well as the expectations of speakers about their listeners. As pointed out earlier, the degree of tangibility of all of these issues varies which can present a serious difficulty when it comes to modelling or controlling them.

Moreover, the assumption that a speaker is capable of producing a near authentic pronunciation of a particular foreign inclusion does not necessarily imply that the listener will actually understand, let alone like and accept what they hear. Wijngaarden

and Steeneken (2000) show with a combination of speech reception threshold (SRT)[3] tests and letter guessing procedure (LGP)[4] experiments that non-authentic, i.e accented pronunciation of foreign words increases intelligibility to inexperienced learners. Experienced L2 learners, on the other hand, perceive non-accented L2 speech as more intelligible. Their experiments involves native Dutch speakers listening to German and English speech produced by German and English native speakers as well as Dutch native speakers. If the same holds true for the perception of polyglot speech, where the foreign items are spoken in the context of the listeners own language, is still an open question.

Even if a TTS synthesis system is capable of producing authentic pronunciations of embedded inclusions or can at least approximate them to some extent in the foreign language, it is unlikely to be the accepted or preferred rendering compared to one that adjusts at least some foreign speech sounds to the acoustic characteristics of the listener's own language. As Lindström and Eklund (2000) point out, too much approximation to the foreign language may not result in the desired effect for listeners. They may consider it exaggerated or inadequate. The other extreme, producing the pronunciation of embedded foreign words merely by means of letter-to-sound rules of the base language of the text, will result in bad mispronunciations or overly accented speech which listeners are likely to deem uneducated or may not even understand.

Besides speaker- and word-related factors such as age, experience of the foreign language, awareness of the task or frequency of the foreign word, which can heavily influence the results of a mixed-lingual speech perception experiment, a perception study of mixed-lingual TTS synthesis output also depends on the actual synthesis approach of foreign inclusions. Although the chosen approach may well be modelled on the behaviour of subjects in production experiments, the perception study subjects

[3] An SRT test involves adding masking noise to test sentences. If subjects listening to the test sentence repeat all its words correctly, the noise is increased, otherwise it is decreased. The SRT value is the average speech-to-noise ratio over a set of sentences and provides a reliable measure for sentence intelligibility in noise.

[4] LGP was devised by Shannon and Weaver (1949) as a way of estimating lower and upper bounds of linguistic information content of a language. Subjects are asked to guess a string of text one letter at a time without receiving any prior information. After each guess, the correct letter is either revealed (single-LGP) or kept undisclosed until the guess is correct (multiple-LGP). Subjects can make use of the letters guessed up to the current point for predicting the next letter. The more letters a subject guesses randomly, the less redundant the language is to them and the more linguistically skilled they are in that language (van Rooij and Plomp, 1991). LGP scores are expressed in terms of linguistic entropy L (in bits): $L = -log_2(c)$ where c is the fraction of correctly guessed letters.

might not actually accept the output produced. Plus, the system design may simplify and only approximate a certain theory and therefore fall short of the pronunciation quality that a human speaker would produce.

Abresch (2007)'s perception study on polyglot speech produced by a human speaker has shown that listeners prefer the original rendering of some foreign phones but also favour substitutions with native phones over others. This knowledge must be taken into consideration when devising a synthesis strategy for foreign inclusions. Perception studies on mixed-lingual speech produced by polyglot TTS synthesis system have not been carried out to date. It is therefore relatively difficult to predict the possible outcome of a task-based TTS synthesis evaluation. Intuitively, it is expected that well educated German speakers would prefer synthesised speech in which English inclusions are produced at least to some extent with an English-like pronunciation. The next section first presents a general overview of the processes involved in a TTS system, secondly examines how TTS synthesis output is evaluated and finally addresses work on some of these individual processes to enable polyglot TTS synthesis.

6.1.2 Brief Overview of a TTS System

A TTS synthesis system is a computerised system which converts arbitrary written text into synthetic speech. Its main aim is to produce intelligible and natural sounding speech. A TTS synthesiser involves three main stages: **text processing** to extract available information from the input text, **prosody generation** to model variations in pitch, loudness and syllable length and **waveform generation** to produce the synthetic speech. Each stage is briefly summarised below based on information published in Dutoit (1997), Holmes and Holmes (2001) and handouts of the speech processing courses held by Dr. Simon King at the University of Edinburgh in 2002.

6.1.2.1 Text Processing

The text processing component consists of several sub-tasks to extract maximum information from the text which in turn serves as input to subsequent components. Firstly, the text is subjected to a pre-processing step where it is tokenised and normalised. The normalisation step generally involves identifying numerals, abbreviations and acronyms and expanding them into their written-out equivalents if necessary.

Moreover, punctuation ambiguity is resolved. The text is then processed by a morphological analyser which decomposes words into their component parts, i.e. roots and affixes. During syntactic analysis, tokens in the input text are assigned POS tags whereby each token is analysed according to its surrounding context to resolve POS ambiguities. Then, the phrase structure of each input sentence is determined to conduct phrase break prediction. While some systems employ a full syntactic parser (see Chapter 5), others simply perform a superficial syntactic analysis for this task.

Subsequently, the pronunciation of individual words in the text and their lexical stress is determined by a combination of lexicon lookup, letter-to-sound rules and post-lexical rules. Each word is looked up in the integrated lexicon, a large database that contains POS information, syllable structure and lexical stress. In case the pronunciation of a word is obtained by combining different morphs in the lexicon, lexical stress rules are employed to determine the stress pattern for the entire word. The letter-to-sound rules are applied whenever the pronunciation of a word cannot be determined by means of the lexicon. This often happens for new and foreign words or names that occur in the text. In such cases, the pronunciation is predicted from the orthographic form alone. Since letter-to-sound rules are designed for the base language of text that is to be synthesised, they are clearly unsuitable for deriving the pronunciation of foreign inclusions. Once the phone sequence is determined, further rules are required to assign the lexical stress. Hand-written post-lexical rules are applied to achieve effects on pronunciation that cannot be determined for words in isolation. Such adjustments include vowel reduction or phrase-final devoicing in English. At the end of this process, a complete phonetic representation of each utterance is obtained.

6.1.2.2 Prosody Generation

The next step is to generate the prosody for each utterance. Many systems firstly locate and identify symbolic prosodic labels and then use this information for determining the appropriate fundamental frequency (F_0) contour and duration. The first step is achieved by assigning pitch accents and boundary tones to the syllables according to the underlying intonation model (e.g. acoustic, perceptual or linguistic). For example, the linguistic ToBI (Tones and Break Indices) intonation theory (Silverman *et al.*, 1992) is based on a discrete set of symbolic accents types and boundary tones and is frequently used as the standard intonation model. ToBI labels are predicted using

automatic data-driven methods, like for example decision tree models such as Classification and Regression Tree (CART) models (see Breiman *et al.* (1984)). Once the abstract labels are assigned to the utterance, they have to be transformed into numerical F_0 targets and converted into a F_0 contour. There are many different approaches to this task, including the computation of an average pitch contour as well as rule-based or statistical methods.

Each segment to be synthesised must also have a particular duration such that the synthesised speech mimics the temporal structure of typical human utterances. The duration of a speech sound may vary depending on a series of factors, including speech rate, stress patterns, their position in the word and in the phrase as well as phone-intrinsic characteristics. Traditionally, a suitable duration for each phone is estimated on the basis of rules. With the availability of labelled speech corpora, data-driven methods are used to derive duration models automatically.

6.1.2.3 Waveform Generation

Finally, the speech waveform is generated. The various types of approaches to waveform generation include rule-based, concatenative and HMM-based synthesis. Rule-based TTS synthesis (e.g. MITalk, Allen *et al.* (1987)) relies on a simplified mathematical model of human speech production. On the other hand, concatenative synthesis systems (e.g. Festival, Black and Taylor (1997)) exploit recorded speech data. Most current commercial TTS systems employ concatenative synthesis, either by diphone or unit selection. For diphone synthesis, all possible diphones (sound-to-sound transitions) in a particular language need to be recorded and labelled. The speech database only contains one example of each diphone spoken by the same speaker. During synthesis, the necessary diphones are concatenated and the target prosody is superimposed by means of digital signal processing techniques like Linear Predictive Coding (LPC, Markel and H. (1976)), Time-Domain Pitch-Synchronous-OverLap-Add (TD-PSOLA, Moulines and Charpentier (1990)) or Multi-Band Resynthesis OverLap-Add (MBROLA, Dutoit and Leich (1993)). Conversely, unit selection synthesis involves less digital signal processing. However, it requires several hours of recorded speech data. Each recorded utterance is segmented into units of various sizes, including phones, syllables, morphemes, words, phrases and sentences. This segmentation is typically performed by means of a speech recogniser and hand correction. Each unit is

then stored in the database according to its segmentation and acoustic features. During synthesis, a search algorithm selects the best sequence of units from the database for concatenation. The search is generally performed by means of decision trees. A further approach to waveform generation is HMM-based synthesis. According to this method, the recorded, segmented and labelled speech data is used for modelling the speech frequency spectrum, F_0 and duration simultaneously by Hidden Markov Models (HMMs) which then generate speech waveforms based on the Maximum Likelihood criterion.

Given the various processes involved in TTS synthesis, it becomes clear that particularly the steps in the front-end of the system are language-dependent. Moreover, current state-of-the-art TTS systems rely on recorded speech data in a particular language. A system that is able to synthesise mixed-lingual input would not only require a text processing step which identifies foreign inclusions but would also necessitate an appropriate grapheme-to-phoneme conversion as well as suitable speech data. After examining different evaluation methods for TTS synthesis, research on polyglot TTS approaching such issues is presented.

6.1.3 Evaluation of TTS Synthesis

There are different ways to evaluate the quality of synthetic speech. This section mainly focuses on two commonly used subjective tests based on listeners' responses, namely **absolute category rating** and **pair comparison**.

6.1.3.1 Absolute Category Rating

Absolute category rating (ACR), also referred to as single stimulus method, is the most common and straightforward method to evaluate synthetic speech quality numerically (Nusbaum *et al.*, 1984). Subjects are asked to rate each test signal once using a five-point scale that ranges from bad (1) to excellent (5) (CCITT, 1989). The average score or mean opinion score (MOS) of each competing TTS system is therefore determined as the arithmetic mean of its individual signal scores. As this is a subjective evaluation, variability between subjects can be high. Evidently, this method becomes more reliable the higher the number of test speech signals and listeners. The use of a set of reference signals in the evaluation can also help to normalise for listener-dependent variations.

Rather than determine the overall speech quality of synthesised speech as per-

ceived by listeners in one score, some categorical rating methods assess synthetic speech according to separate aspects such as pronunciation, speed, distinctness, naturalness, stress, intelligibility, comprehensibility or pleasantness. This evaluation approach highlights individual strengths and weaknesses of a system and is also easy to set up. The results enable an interpretation in terms of the quality of each system and also give an idea of how different systems compare.

6.1.3.2 Pair Comparison

The acceptance of synthesised output is generally determined by means of a pair comparison (PC) test (Björkman and Gösta, 1957). Listeners are sequentially presented with the same synthetic speech data produced by different systems and have to specify their preference for each pair. This is a forced choice evaluation, i.e. two stimuli cannot be judged equal (Goldstein, 1995). The various synthesis versions have to be presented in all possible orders in order to avoid confusion from order effects. This is also a relatively easy evaluation method to implement. PC gives evidence as to which system is the preferred one but does not indicate the actual quality of each system.

6.1.4 Polyglot TTS Synthesis

Traditionally, TTS synthesis systems are designed to process monolingual texts. When encountering multilingual texts, one synthesis strategy is to change the voice at every language change. This is suitable for documents containing paragraphs in different languages. Once the various languages of individual sections have been detected, the system switches to the corresponding voice. Multilingual TTS synthesis with different voices as implemented by Turunen and Hakulinen (2000) for Finnish and English lends itself well to processing, for example, multilingual email messages containing translations or sections in different languages. However, as seen throughout this thesis, language changes can happen on much lower levels, including the sentence or even the individual word or sub-word level. Turunen and Hakulinen (2000) point out that word-level language changes are actually the most common form of multilingual e-mail content. Running a multilingual synthesiser over such mixed-lingual text would result in frequent voice changes which tend to irritate users even if they do not have an effect on their comprehension. However, Turunen and Hakulinen (2000) claim that

changing to a voice with opposite gender from the previous one, as performed in their system, helps users to cope with voice changes better. This may hold true for language changes on the sentence level or higher but may be more confusing than beneficial if voice and voice gender changes happen within sentences.

Thus, the multilingual TTS synthesis paradigm is unsuitable for truly mixed-lingual texts where language changes occur within sentences and phrases. Optimal naturalness and intelligibility of synthesised mixed-lingual speech would be obtained if the voice does not change but adapts its pronunciation to the foreign language whenever it encounters an embedded foreign inclusion. This type of TTS system, i.e. one that is able to synthesise several languages using the same voice with appropriate pronunciation is referred to as **polyglot synthesis** (Traber *et al.*, 1999). In recent years, polyglot synthesis has been addressed by a series of research efforts on various sub-tasks of a TTS system to enable processing of mixed-lingual data: generating **multi-context rules** for phonological processing (e.g. Romsdorfer and Pfister, 2004), devising **prosody control** for polyglot TTS (e.g. Romsdorfer *et al.*, 2005), constructing **multilingual vocalic databases** (e.g. Traber *et al.*, 1999), developing **phoneme mapping** (e.g. Campbell, 1998; Badino *et al.*, 2004) and creating an **average polyglot voice** (e.g. Latorre *et al.*, 2005) by combining voices of monolingual speakers in different languages.

6.1.5 Strategy for Task-based Evaluation with TTS

Section 6.1.1 examined production and perception studies carried out for mixed-lingual text as well as in the field of SLA. One reoccurring conclusion is that, unless particularly fluent in a foreign language, speakers tend to adapt articulatory and phonetic characteristics of many foreign speech sounds to those of similar sounds in their native tongue which results in a foreign accent. However, there are certain new speech sounds in a foreign language called xenophones which do not have a close equivalent representation in the native language and can still be produced more or less successfully by a majority of people.

This phenomenon has been modelled to varying degrees by research efforts on polyglot TTS. The synthesis of mixed-lingual text by maintaining the same voice identity is the common goal of a series of methods, including the creation of multi-

lingual vocalic databases, phone mapping algorithms and the building of an average voice by means of HMMs. The underlying goal of the first method is to achieve near perfect pronunciation of the foreign language inclusions. Although speakers are unlikely to pronounce foreign words perfectly, such a system would be very useful for multilingual TTS synthesis, for example to synthesise various paragraphs in different languages with the same voice as could be expected by an E-mail client of an international organisation. Another application area of this method could be foreign language teaching. However, the approach of building a multilingual voice database by means of a polyglot speaker is language-dependent and therefore difficult to extend to new languages. In comparison, the second method, phone mapping, is less dependent on the availability of speech resources and polyglot voice talents and therefore easier to expand to new language scenarios. Moreover, it is based on the assumption of approximating speech sounds which is a prevalent phenomenon in the pronunciation of foreign words. Finally, the development of an average voice is the most language independent approach out of all three and therefore most easy to extend. However, the fact that is not as high quality as unit selection synthesis could have serious effects on the perception of a synthesised utterance in general and make differences in the synthesis of foreign inclusions less obvious to the listeners. Unfortunately, these methods have not been evaluated on real mixed-lingual utterances. Considering the results of production and perception studies of polyglot speech reviewed in Section 6.1.1.1, the phone mapping method appears as the most promising one with results that are likely to be most accepted in a subjective perception study. A readily available system set up for mixed-lingual synthesis would be ideal for a task-based TTS evaluation of the English inclusion classifier. Researchers at Nuance and Loquendo S.p.A. approach polyglot synthesis via multilingual vocalic databases and phone mapping, respectively. Either system would be interesting to test in a perception experiment. If such a system is not available, a voice needs to be created and the phoneme mapping algorithm implemented in order to be able to synthesise foreign inclusions. This can be facilitated by the speech synthesis system Festival[5] (Black and Taylor, 1997) and the FestVox[6] voice building tools.

[5] `http://www.cstr.ed.ac.uk/projects`
[6] `http://www.festvox.org`

6.1.5.1 Polyglot TTS system

The main aim is to carry out synthesis experiments on German utterances containing English inclusions as a task-based evaluation of the English inclusion classifier introduced in this thesis. As previously mentioned, one version of the Loquendo TTS system is specifically designed for mixed-lingual input. The system is able to alternate between languages via the control tag ***\lang=<language>*** in the markup of the input as shown in the example below. ***\lang=*** is used to return to the native language of the voice.[7]

> *\lang=English* Security-Tool *\lang=* verhindert, dass *\lang=English* Hacker *\lang=* über *\lang=English* Google *\lang=* Sicherheitslücken finden
>
> Translation: *Security Tool prevents hackers from finding security holes via Google.*

Such a setup would allow for a perception experiment comparing an all German baseline synthesiser with one that is able to deal with foreign language inclusions. Three types of synthetic stimuli can be produced when using mixed-lingual input: (1) stimuli without markup for synthesis with the German baseline system, (2) stimuli marked up correctly for English inclusions for synthesis with a polyglot TTS system and (3) stimuli marked up by the English inclusion classifier for synthesis with a polyglot TTS system. The second type of stimuli will provide a clearer idea of whether the chosen polyglot TTS synthesis paradigm improves the output quality of mixed-lingual data. With this in mind, the third type of stimuli can then be used for the actual task-based evaluation of the foreign inclusion classifier.

6.1.5.2 Experimental Setup and Evaluation

Before proposing exact details of the evaluation procedure, it should be reiterated that all factors learnt from previously described production and perception studies of mixed-lingual speech are likely to affect the results of the perception study of mixed-lingual TTS synthesis (see Section 6.1.1.1). Regarding subject-specific characteristics, age, level of English language skills and educational background must be controlled

[7] A demo of this system is available online at: `http://actor.loquendo.com/actordemo/ml.asp?language=en`

for. Ideally, the subject group should be made up of well-educated German subjects of different age groups with at least some basic knowledge of English. However, we have also learnt that social and psychological factors can play a role in how embedded foreign words are perceived. It may be more difficult to control these as they are highly dependent on subjects' individual personalities and preferences. Concerning stimuli-related factors, the type, length, frequency, and age of an English inclusion should be taken into account. Another issue that must be considered is the context in which inclusions appear. Considering the findings made by Abresch (2007), it is not effective to evaluate the synthesis of actual foreign inclusions themselves as it may influence subjects' perception and make them aware of the purpose of the experiment. Moreover, subjects may prefer a more authentic rendering of the English inclusions out of context than if they are embedded in a German sentence. Consequently, choices have to be made with regard to selecting German carrier sentences containing English single- and/or multi-word inclusions for synthesis. It would be interesting to use the carrier sentences used in the perception study of Abresch (2007) as that would allow parallels to be drawn with her results for human rather than synthesised speech. Some more thought has to go into the actual selection of utterances, as a task-based evaluation of the English inclusion classifier will be part of this study.

For the evaluation, subjects should be asked to evaluate the synthetic speech in terms of intelligibility and naturalness. There are various suitable evaluation techniques for determining these features. As described in Section 6.1.3, two commonly used and straightforward methods are ACR or PC. As the results of ACR enable an interpretation in terms of the quality of each system and also give an idea of how different systems compare, this evaluation method lends itself well to the proposed evaluation.

6.2 Machine Translation

A further application for which the detection of English inclusions is expected to be beneficial is machine translation (MT). As foreign inclusions carry critical content, their correct detection will provide vital additional knowledge to MT systems. The occurrence of English inclusions in other languages is a non-negligible phenomenon. This has been particularly apparent during my work as a translation quality consultant for Verbalis Ltd., a Scottish start-up company which provides high-speed language

translation with an example- and analogy-based MT system. English inclusions are extremely frequent in documents of global software or finance companies which translate all of their publications into other languages.

An MT-based integration and evaluation of the English inclusion classifier is, however, not trivial, as such language mixing requires different translation strategies depending on the type of inclusion. For example, foreign proper nouns appearing in the source text mostly do not require translating in the target language. However, a translation is likely to be preferred for common nouns if the target language is not English. This is illustrated in the following German sentence and its French, English and Russian translations. The English common noun *Crew* is translated into the French noun *équipage*, the Russian noun *экипаж*[8], or decapitalised when translated into English. Conversely, the English proper name *Endeavour* is used in all three languages in Latin script.

GERMAN: Die Endeavour-Crew blieb elf Tage im All.
FRENCH: L'équipage d'Endeavour est restée onze jours dans l'espace.
ENGLISH: The Endeavour crew stayed in space for eleven days.
RUSSIAN: Экипаж Endeavour провел одиннадцатьдней в космосе.

As many proper nouns are treated as unseen words by MT engines, they are transferred to the target language as they are. English common nouns appearing in German text, on the other hand, often require translating particularly when the target language is not English. The latter is illustrated in the following German sentence containing the two English nouns *Tools* and *News*. The English and French MT output of this sentence as produced by BABELFISH[9], a rule-based MT system, is presented below. The system evidently treats both inclusions as unseen words and simply reinserts them into the output sentence. Interestingly, one of the English nouns (*news*) is decapitalised in the German-English translation.

[8] Interestingly, the Russian translation of the English noun *crew* is a borrowing of French origin.
[9] `http://babelfish.altavista.com`

GERMAN:	Mit diesem **Tool** können Sie parallel in sämtlichen **News** suchen.
ENGLISH(BABELFISH):	With this **Tool** you can search parallel in all **news**.
FRENCH(BABELFISH):	Avec ce **Tool**, vous pouvez chercher parallèlement tous les **News** dans.

While it was established in Sections 2.1.3.2 and 4.2 that French contains a large number of anglicisms at least in the domain of IT, the use of such anglicisms may, however, not always be the preferred choice by a human translator (HT). They may produce the following French translation of the German sentence:

FRENCH(HT): Avec cet **outil**, vous pouvez chercher tous les **actualités** en parallèle.

Mixed-lingual compounds or interlingual homographs are even more of a challenge to MT systems. One very interesting example occurs in the following German query:[10]

GERMAN:	Nenne einen Grund für Selbstmord bei **Teenagern**.
ENGLISH(BABELFISH):	Call a reason for suicide with **dte rodents**.
FRENCH(BABELFISH):	cite une raison de suicide avec des **Teenagern**.

The English inclusion *Teenager* appears in the dative plural and consequently receives the German inflection *n*. Instead of treating this noun as an English inclusion when translating the sentence into English, Babelfish processes this token as the German compound *Tee+Nagern* (tea + rodents) and translates its subparts into the token *dte*[11] and the noun *rodents*. Translating into French, the MT system treats the English inclusion as unseen and inserts it directly into the translation without further processing, such as inflection removal. Combined with a named entity recogniser, the English inclusion classifier could signal to the MT engine which items require either translating or transferring with respect to the target language.

Multi-word English inclusions also pose difficulty to most MT systems. If the system has not encountered a particular expression in its training data or its lexicon, it is likely to treat the entire expression as unseen. However, MT systems are not necessarily aware of the boundaries of such multi-word expression as illustrated in

[10]This query appeared in CLEF 2004 where one of the tasks was to find answers to German question in English documents (Ahn *et al.*, 2004).

[11]Note that *dte* is not a typo but an error in the MT output.

the following German example sentence, its English gloss and translations by ten MT system demos currently available online.

GERMAN:	**Made in Germany** ist gefragt.	
ENGLISH(HT):	**Made in Germany** is in demand.	
ENGLISH(LOCAL TRANSLATION):[10]	Larva in Germany is in demand.	✗
ENGLISH(COMPENDRIUM):[11]	Maggot in Germany is in demand.	✗
ENGLISH(FREETRANSLATION.COM):[12]	Maggot in Germany is asked	✗
ENGLISH(HYPERTRANS):[13]	Grub in Germany is asked.	✗
ENGLISH(INTERTRAN):[14]	maggot in Teuton am asked.	✗
ENGLISH(PERSONAL TRANSLATOR):[15]	Maggot in Germany is in demand.	✗
ENGLISH(POWER TRANSLATOR):[16]	**Made in Germany** is in demand.	✓
ENGLISH(REVERSO):[17]	Maggot in Germany is asked.	✗
ENGLISH(SYSTRAN):[18]	Larva in Germany is in demand.	✗
ENGLISH(@PROMT):[19]	Maggot in Germany is asked.	✗

The translations by the different MT engines show that most systems (9/10) struggled to transfer the multi-word English inclusion back into English. Only one system (POWER TRANSLATOR) produced a correct translation, identical to the human rendering. When surrounding the English inclusion by quotation marks in the German input sentence, one other system (HYPERTRANS) was able to translate the expression correctly. All other systems produced the same wrong translation as already listed. It can be seen that the majority of systems (5/9), that did not produce a correct translation, managed to process the German portion of the sentence correctly but failed on the English inclusion. Most systems failed to recognise the entire English phrase as unseen. Instead of transferring it, they mistook *Made*, the first word of the inclusion, as the German noun *Made* (maggot, larva or grub). Only one system (INTERTRAN)

[10] http://www.localtranslation.com/
[11] http://www.translendium.com/
[12] http://www.freetranslation.com/
[13] http://www.dagostini.it/hypertrans/index.php
[14] http://www.tranexp.com:2000/Translate/result.shtml
[15] http://www.linguatec.net/onlineservices/pt
[16] http://www.lec.com/w2/translate-demos.asp
[17] http://www.reverso.net/text_translation.asp?lang=EN
[18] http://www.systran.co.uk/
[19] http://www.e-promt.com/

attempted to translate the prepositional phrase *in Germany* from what it thought to be German text into English as *in Teuton*. Bearing in mind that this is only one example, it only allows for limited conclusions to be drawn. However, this example clearly demonstrates that English inclusion detection can be beneficial to MT systems, particularly for multi-word expressions. A large-scale experiment would need to be carried out in order to quantify this claim.

The English inclusion classifier performs language classification on the token level. However, in Chapter 5 on parsing sentences containing English inclusions, it was shown that contiguous tokens of English origin tend to belong to the same phrase constituent. Consequently, the English inclusion classifier can be used to identify both single- and multi-token inclusions. This information can then be used by the MT system in order to determine whether an inclusion is unseen or contained in the knowledge base of the particular engine. Applied in this way as a pre-processing step, English inclusion detection would be particularly useful to rule-based MT systems. However, it is anticipated that even example-based and statistical MT systems which rely on either a translation memory of example translations or a parallel corpus would benefit from English inclusion detection particularly for unseen expressions. In future, it might be possible to grant current bi-directional MT systems access to knowledge bases from other languages along the same lines as polyglot TTS. Whenever, the MT system encounters an expression of a different language origin than the base language of the text that is to be translated, it will be directed to access lexicons and corpora in the language identified by the inclusion classifier.

Some MT systems are already deliberately designed to process certain expressions differently from the remainder of the text. For example, Koehn (2004) has developed a freely available beam search decoder for phrase-based statistical MT models to determine the translation probability for translating a source language sentence (s_{SL}) into a target language sentence (s_{TL}). This is done by means of a translation model $p(s_{SL}|s_{TL})$, and a target language model $p(s_{TL})$ as follows:

$$argmax_{s_{TL}} p(s_{TL}|s_{SL}) = argmax_{s_{TL}} p(s_{SL}|s_{TL}) p(s_{TL}) \tag{6.1}$$

The decoder is designed to process external knowledge specified in XML markup

which can contain one or multiple target language translations with or without probabilities assigned to them (see Example below). The XML attributes essentially replace the translation options from the phrase translation table. If only one translation is specified, the probability for that phrase is 1. However, multiple options can be specified along with their translation probabilities, and the target language model is used to determine which option is the preferred one. After identifying the instances of English inclusions that are likely to be translated into the target language, their possible translations can be determined by means of a translation dictionary or a parallel corpus. The translation options can subsequently be presented to the decoder as external knowledge in the XML markup and ranked by the target language model to determine the overall best translation of the sentence.

```
Die <n french="équipage">Crew</n> der <ne
french="Endeavour">
Endeavour</ne> blieb elf Tage im All.
```

Translation: *The Endeavour crew stayed in space for eleven days.*

For such an experiment, a parallel source and target language corpus is required in order to construct both the translation model and the target language model. A parallel English-target language corpus is also necessary to determine possible translation options for English inclusions. The benefit of providing additional information on English inclusions, combined with a named entity recogniser, can be determined by running the MT system over German text with and without marked inclusions. The output quality can then be evaluated by means of evaluation metrics commonly applied for statistical MT. These include human judgements with regard to the syntactic and semantic well-formedness of a sentence or automatic evaluation in terms of word error rate or BLEU score (Papineni *et al.*, 2001) in relation to reference translations.

6.3 Linguistics and Lexicography

Finally, the English inclusion classifier presents a useful tool for linguists and lexicographers whose job it is to study languages and compile lexicons and dictionaries. Many linguists are interested in language mixing and study the influence that English is having on other languages. The rise in the number of publications and entire conferences and workshops dedicated to the occurrence of anglicisms and foreign inclusions

in other languages (e.g. the international conference on *Anglicisms in Europe* (2006) and the workshop on *Strategies of Integrating and Isolating Non-native Entities and Structures* to be held in February 2008) illustrate this trend. All of the studies on the frequency of anglicisms and English proper names in German discussed in Section 2.1.3.2 are based on painstaking manual analysis and frequency counting in corpora. For example, Furiassi (2007) reports on work carried out by one of his students who manually examined over 28,000 concordances to identify anglicisms in Italian. While such analysis cannot be completely automated, the output of the English inclusion classifier can be used to assist in identifying anglicisms and therefore alleviate the burden for linguists.

Some preliminary attempts in automating the detection of anglicisms have been made by Furiassi and Hofland (2007). However, they do not evaluate the performance of their character-based n-gram algorithm. Moreover, previous work on LID reviewed in Section 2.2 shows that character-based n-gram algorithms work well on large passages of text but not on single tokens. Furiassi and Hofland (2007) also attempted to classify specifically false anglicisms, i.e. pseudo-anglicisms. It is unclear how their algorithm can differentiate between false and real anglicisms. Both types are made up of English morphemes and the difference between them is often a semantic one. Identifying all types of anglicisms with English forms, following Onysko's model (see Section 2.1.2.4), may be a more realistic aim. This was attempted by Bartsch and Siegrist (2002) who derived a list of typical English morphological endings and character sequences from the Porter stemmer (Porter, 1980) in order to identify anglicisms in the *Darmstadt Corpus of German Languages for Specific Purposes*. This semi-automatic algorithm is also not evaluated but it is expected that many anglicisms are missed with this technique. Nevertheless, it is clear that any type of assisted language analysis will support linguists who are otherwise required to examine large amounts of text manually or revert to drawing conclusions based merely on estimations.

In future work, it would be interesting to run the English inclusion classifier over articles published in the German magazine *Der Spiegel* and compare the results with the manually guided analysis made by Onysko (2007). This will make it possible to compare how different the results are and, in turn, lead to further adaptation of the classifier. Furthermore, the English inclusion classifier can be used for diachronic analysis of language in order to determine the frequency of anglicisms over a given

time period. For the purpose of linguistic analysis, this will reveal diachronic language changes. For example, it can be investigated if the use of specific inclusions has increased or decreased and how this relates to the use of their German equivalents. Given meta data information on the subject matter of an article, analysis could be limited to certain domains instead of all documents. Combined with a dictionary of anglicisms, it will also be possible to identify the appearance of new anglicisms. This will allow us to draw conclusions about the influence of English on German over time and highlight borrowing trends. Moreover, it will be possible to draw conclusions with regards to localised borrowing, i.e. if the use of certain inclusions is limited to specific documents. For example, a journalist may use an English expression once because it was relevant to the topic of the specific article but would not use that expression in other contexts.

Determining the frequency of anglicisms in a given corpus is not only useful when examining language developments. It can also assist lexicographers when deciding which new words to add to a dictionary or lexicon. The general usefulness of NLP tools to lexicography is addressed, for example, by Kilgarriff (2003; 2005) who believes that lexicographers are best supported by linguistically-aware corpus query tools. Kilgarriff (1997) stresses the importance of frequency information for all entries in a dictionary for language learners based on the assumption that it is more important to learn common terms than uncommon ones. There are several dictionaries that provide this kind of detail, e.g. the Collins COBUILD English Dictionary (1995), the Longman Dictionary of Contemporary English (1995) or the Russian Learner's Dictionary: 10,000 Words in Frequency Order (1996). Regarding the extension of dictionaries and lexicons with neologisms, Breen (2005) reports on a strategy of combining a parser with lexicon lookup to harvest a list of unknown katakana words in Japanese. This candidate list can then be manually checked by the lexicographer to determine which words should be entered into the lexicon.

It emerged in an interview with lexicographers at Chambers Harrap Publishers Ltd. in Edinburgh conducted in June 2006 that there is no scientific mechanism for lexicographers to decide on when to add a new word into the lexicon or dictionary. When loan words are used extensively in a language, sometimes up to the point where they are no longer perceived as foreign, they tend to be added. This decision is relatively arbitrary and mostly down to the individual lexicographer after having come across a certain term for a number of times. Furiassi (2007), for example, calls for a

clear strategy for making this decision. He argues that non-adapted anglicisms should be entered into general dictionaries or lexicons if they occur above a certain frequency in a large and balanced corpus. The English inclusion classifier would be a useful tool in this context. It could be constantly run over new documents, thereby allowing lexicographers to identify new loan words, possibly even trace them, and determine the frequency of a certain loan word over time. The English inclusion classifier can consequently make lexicographers aware of a language mixing phenomenon that they might otherwise miss during their corpus analysis. Equally, lexicographers could feed their knowledge back into the classifier as a way of improving its performance. In this way, the classifier would allow lexicographers to base their decisions to include a term in the dictionary based on empirical facts, and, conversely, the lexicographers' knowledge could be exploited to increase the performance of the classifier.

6.4 Chapter Summary

This chapter described in detail the usefulness of English inclusion detection for various applications and fields, including TTS, MT and linguistics and lexicography. As with parsing, input to TTS and MT systems is generally assumed to be monolingual and so far there has been little focus on devising systems that are able to process mixed-lingual input sentences. In our increasingly globalised world where English is infiltrating many other languages, automatic natural language processing must be able to deal with such language mixing. The English inclusion classifier could be used in a pre-processing stage in order to signal where language changes occur. Further processing of English inclusions then depends on various synthesis or translation strategies for specific cases. This chapter reviewed previous work on deriving such strategies and presented some ideas for future work in terms of extrinsically evaluating the benefit of English inclusion detection for both applications.

Regarding the fields of linguistics and lexicography, this chapter summarised the benefits of the English inclusion classifier as a tool for automating synchronic and diachronic language analysis. As such, the classifier could be beneficial to linguists who examine the frequency of certain expressions at a given point in time, or in different domains, and who track language changes over time. Moreover, it could be used to assist lexicographers in their decisions to include specific terms into lexicons or dictionaries.

Chapter 7

Conclusions and Future Work

This thesis has shown that it is possible to create a self-evolving system that automatically detects English inclusions in other languages with minimal linguistic expert knowledge and no ongoing maintenance. This **English inclusion classifier** has shown three key advantages in that it is **annotation-free**, **dynamic** and **easily extensible**.

The fact that the English inclusion classifier is **annotation-free** represents an advance over existing statistical-based NLP systems which require annotated training data, a dependency that is referred to as the annotation bottleneck. When applied to a new problem or domain, statistical systems will fail without this annotation. This weakness has been demonstrated here with an experiment that applied a machine learning approach to English inclusion detection. A further experiment with the machine learner also determined that an annotated data pool of over 80,000 tokens is required to reach even a comparable performance to the English inclusion classifier developed here. In fact, the classifier does not require any overhead in terms of extensive, and consequently expensive, manual annotation when introduced to a new domain. Therefore the English inclusion classifier is readily applicable to unseen data sets and has been experimentally shown to perform well under these circumstances.

The English inclusion classifier is **dynamic** because of its search engine component. As the Internet provides access to extremely large quantities of evolving data in different languages, search engines can be used to determine the estimated relative token frequencies for new and unseen words. This classifier therefore exploits the volume of data published online to perform mixed-lingual language identification. The thesis has also presented a corpus search experiment with various sizes of corpora

that quantified the advantage of access to the Internet over a currently available large corpus. The results of this experiment highlight the merits of this novel approach to mixed-lingual language identification.

Finally, the classifier is **easily extensible** to any new language using Latin script. This aspect is critically important as we have seen that language mixing affects many different languages. It has been shown that given a POS-tagger and a lexicon, the system can be extended to a new language within one person week. Only minimal language-specific knowledge is required to derive a set of post-processing rules in order to resolve any ambiguous cases, thereafter a good performance gain can be rapidly achieved. Such swift extensibility in language specific classification is unusual, and is a key advantage of the system developed here.

All three of these advantages combine to create a self-evolving, scalable and adaptable system that automatically detects English inclusions in other languages with minimal human interaction or ongoing maintenance. In fact, it was demonstrated in this thesis that interfacing the English inclusion classifier with two German parsers can improve the quality of their output.

7.1 Thesis Contributions

The main contributions of this work to English inclusion classification are:

- The development of an annotation-free, dynamic and extensible English inclusion classifier for German or French.
- Extensive evaluation of classifier performance on different languages, domains and data sets.
- Interfacing of the new English inclusion classifier with two German parsers specifically to improve their performance.
- The preparation of annotated German and French gold standard corpora for English inclusions. These will be released to enable comparison of this work with any that may be done by other research groups in future.

- Identification of further applications for which English inclusion detection would be useful. This includes the outlining of extrinsic evaluations and other experiments in TTS, MT, and linguistics and lexicography.

7.2 Future Work

There are three clear paths for future research that significantly extend the functionality and usability of the classifier developed in this thesis.

The first path would be to develop a new algorithm for processing non-Latin script. This algorithm would need to be capable of extracting both phonetically transliterated and directly included words in languages such as Russian or Arabic. This is a significant challenge because access to transcribed English corpora in different languages is minimal at best.

The second path is to extend the classifier to recognise language changes at the morpheme level, including mixed-lingual compounds and English inclusions with inflections. To do this, a new layer of processing needs to be added to the system that morphologically analyses each word. The point here is that without consideration of mixed lingual language identification, the NLP community will continue to face significant problems as the phenomenon of language mixing grows.

The third and final path is to evaluate the English inclusion classifier with respect to other potential applications described in Chapter 6. All positive findings from such extrinsic evaluation would make the classifier an attractive tool to be used in the various research fields.

Appendix A

Evaluation Metrics and Notation

This appendix explains the evaluation metrics and statistical significance tests used for the experiments presented in Chapters 3, 4 and 5 and explains how they are calculated. It also specifies various notations in order to avoid confusion.

A.1 System Evaluation Metrics

In the broad sense, English inclusion detection can be regarded as an information extraction task, where the aim is to identify all English inclusions occurring in text that is written primarily in a different language. The English inclusion classifier's performance is evaluated intrinsically on seen and unseen evaluation data against gold standard annotation. The evaluation measures used for this intrinsic evaluation are **accuracy** and **F-score** which are calculated using the `conlleval` script written by Erik Tjong Kim Sang.[1] The identification of English inclusions is therefore evaluated in a similar way to named entity recognition (NER), but for single tokens. A useful way to illustrate how accuracy and F-score are computed is via a contingency table of the gold standard annotation and the system output (see Table A.1). The positive and negative annotations of the gold standard are compared against those produced by the system. The positive and negative labels which are correctly predicted by the system with respect to the gold standard are called **true positives** (TP) and **true negatives** (TN), respectively. A wrongly predicted positive label is called **false positive** (FP) and

[1]This script is freely available at: `http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt`

a wrongly predicted negative one is called **false negative** (FN).

		System output		
	Labels	Positive	Negative	Total
Gold standard	Positive	TP	FN	Gold P
	Negative	FP	TN	Gold N
	Total	Predicted P	Predicted N	All labels

Table A.1: Contingency table of gold standard and system output.

The metric accuracy (*Acc*) represents the percentage of all correctly predicted labels, both positive and negative ones, and **word error rate** (WER) is the percentage of all incorrectly predicted labels.

$$Acc = \frac{\text{TP} + \text{TN}}{\text{All labels}} \tag{A.1}$$

$$WER = \frac{\text{FP} + \text{FN}}{\text{All labels}} \tag{A.2}$$

Balanced F-score (F) represents the performance for positive labels specifically and is calculated as the harmonic mean of **precision** (P) and **recall** (R) as:

$$F = \frac{2 * P * R}{P + R} \tag{A.3}$$

whereby P and R are calculated as follows:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{A.4}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{A.5}$$

In this thesis, all accuracy, word error rate, precision and recall values are multiplied by 100 in order to represent them as percentages. Consequently, they, as well as F-scores, range between 0 and 100.

		Annotator A		
	Labels	Positive	Negative	Total
Annotator B	Positive	$p_A p_B$	$n_A p_B$	p_B
	Negative	$p_A n_B$	$n_A n_B$	n_B
	Total	p_A	n_A	$p+n$

Table A.2: Contingency table of two annotators.

A.2 Inter-annotator Agreement Metrics

Inter-annotator agreement (IAA) is calculated by means of a contingency table for the data versions produced by different annotators, e.g. annotators A and B (see Table A.2). This matrix is essentially the same as the one presented in Figure A.1, only that one annotator represents the gold standard and the other the system output.

Based on counts in the contingency table, IAA scores can then be computed as several metrics: **pairwise accuracy** (Acc) and **F-score**, or the **Kappa coefficient** (κ).

A.2.1 Pairwise accuracy and F-score

Brants (2000a), for example, reports IAA for part-of-speech annotation of the German NEGRA treebank in terms of accuracy and F-score. These are also standard metrics used for NER, whereby accuracy is determined on the token level and F-score on the phrasal level. As the English inclusion annotation was carried out on the token level, the IAA F-score is determined on the token level as well:

$$Acc = \frac{p_A p_B + n_A n_B}{p_A + n_A} = \frac{p_A p_B + n_A n_B}{p_B + n_B} \tag{A.6}$$

$$F = \frac{2 * P * R}{P + R} = \frac{2 * \frac{p_A p_B}{p_A} * \frac{p_A p_B}{p_B}}{\frac{p_A p_B}{p_A} + \frac{p_A p_B}{p_B}} \tag{A.7}$$

As seen in both equations, accuracy and F-score are symmetric between the test and gold data. Accuracy is symmetric, as it is defined as the ratio of the number of tokens on which both annotators agreed over the total number of tokens. F-score is

symmetric, as $precision(A,B) = recall(B,A)$ and balanced F-score is the harmonic mean of recall and precision (Brants, 2000a). The annotations of one annotator (A or B) can therefore be arbitrarily chosen as the gold standard reference.

A.2.2 Kappa Coefficient

While pairwise accuracy and F-score are satisfactory IAA measures, they do not allow a comparison of observed agreement and agreement that occurs completely by chance. An IAA metric that captures this kind of agreement is the kappa coefficient (κ) (Cohen, 1960). The kappa coefficient is commonly used to determine the IAA of corpus annotations (e.g. Carletta, 1996). It measures the observed agreement between two annotators (p_o) taking into account agreement that occurs by chance alone, also called the expected agreement (p_e):

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{A.8}$$

The observed agreement (p_o), which is essentially accuracy, and the expected agreement (p_e) are calculated as follows:

$$p_o = \frac{p_A p_B + n_A n_B}{p_A + n_A} = \frac{p_A p_B + n_A n_B}{p_B + n_B} \tag{A.9}$$

$$p_e = \frac{p_A}{p+n} * \frac{p_B}{p+n} + \frac{n_A}{p+n} * \frac{n_B}{p+n} \tag{A.10}$$

κ-coefficient	Strength of agreement
< 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Table A.3: Agreement interpretation of κ-values (Landis and Koch, 1977).

κ-values can range from -1 (perfect disagreement) to +1 (perfect agreement); a value of 0 corresponds to chance agreement. Although there are no agreed standards, the scale suggested by Landis and Koch (1977) which is shown in Table A.3 is often used for interpreting κ-values.

A.3 Parsing Evaluation Metrics

The English inclusion classifier is also evaluated extrinsically in a series of parsing experiments with a statistical and a rule-based parser (Chapter 5). The performance of the different statistical parsing models, described in Chapter 5.3, are evaluated in terms of a series of metrics including labelled precision, recall and F-score, unlabelled dependency accuracy, and bracketing scores. They are explained in detail below. The output of the rule-based parser, used in the experiments discussed in Chapter 5.4, is merely evaluated in terms of coverage and average number of derivations per sentence.

A.3.1 Labelled Precision, Recall and F-score

Labelled precision, recall and F-score are calculated in the same way as in described in Chapter A.1 but on labelled brackets instead of language identification tags. This means that:

- Labelled precision represents the ratio of the number of correctly labelled constituents in the parse tree and all constituents in the parse tree. A constituent counts as correct if it spans the same words and has the same label as a constituent in the gold tree.
- Labelled recall is the ratio of the number of correctly labelled constituents in the parse tree and all constituents in the gold tree.
- F-score is the harmonic mean of precision and recall.

A.3.2 Dependency Accuracy

Dependency-based evaluation of parsing output was first introduced by Lin (1995) who pointed out that the values of the previously described evaluation metrics can considerably deteriorate in case of a single attachment error that may not be that dramatic

from a linguistic point of view. This is the motivation behind **unlabelled dependency accuracy** (Dep), the evaluation metric proposed by Lin (1995), which is based on comparing dependency tuples in parse and gold trees instead of labelled constituents and phrase boundaries. A sentence is represented in terms of a dependency tree where each word (apart from the head word of the sentence) is the modifier ($M \rightarrow D$) of another word (its head, or H) based on a grammatical relationship. Therefore, each fully parsed sentence and its gold standard tree are made up of $N-1$ dependency tuples, where N is the number of words in the sentence. Dependency accuracy is calculated based on the number of dependents in the sentence that are assigned the same head as in the gold standard ($H(M \rightarrow D)_{correct}$) as:

$$Dep = \frac{H(M \rightarrow D)_{correct}}{N-1} \tag{A.11}$$

It is unlabelled, as the type of relation between the modifier and its head is not considered during evaluation. In order to perform dependency-based evaluation, the constituency trees that are output by the statistical parser must be converted into dependency trees. The conversion algorithm for this procedure is described in detail in Lin (1995).

A.3.3 Bracketing Scores

Parsing performance is also evaluated in terms of **average crossing brackets** (AvgCB), **zero crossing brackets** (0CB) and **two or less crossing brackets** ($\leq$2CB). AvgCB is the average number of constituents in a parse tree that cross the constituent boundaries of the gold tree, e.g. $((W_1\ W_2)\ W_3)$ versus $(W_1\ (W_2\ W_3))$. 0CB is the percentage of sentences for which constituents are non-crossing and $\leq$2CB is the proportion of sentences whose constituents cross twice or less with those of the gold parse tree.

A.4 Statistical Tests

When comparing the performance of the English inclusion classifier to that of another system, or to the baseline, **Pearson's chi-square** (χ^2) test is used for determining statistical significance/insignificance in the difference.

Variable	Correct	Incorrect	Total
Baseline	a	b	a+b
Classifier	c	d	c+d
Total	a+c	b+d	a+b+c+d=N

Table A.4: Contingency table of baseline and classifier in terms of their correct and incorrect labels.

A.4.1 Chi-Square Test

Pearson's χ^2 test, a non-parametric test, is used to determine if the difference in accuracy of the English inclusion classifier and the baseline (or that of another classifier) is statistically significant, the alternative hypothesis (H_1). The hypothesis that there is no significant difference in performance is called the null hypothesis (H_0). If the null hypothesis is rejected, then the difference in performance is regarded as significant, otherwise it is insignificant.

χ^2 is calculated based on the observed frequency (O_i) and the expected frequency (E_i) as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{A.12}$$

O_i and E_i are determined using the variables in the 2x2 contingency table, which refer to the number of correct and incorrect labels of the baseline and a classifier for example (see Table A.4):

$$O_1 = a, \quad E_1 = \frac{(a+b)*(a+c)}{N} \quad \text{and} \quad \frac{(O_1-E_1)^2}{E_1} = \frac{(a-\frac{(a+b)*(a+c)}{N})^2}{\frac{(a+b)*(a+c)}{N}}$$

$$O_2 = b, \quad E_2 = \frac{(a+b)*(b+d)}{N} \quad \text{and} \quad \frac{(O_1-E_1)^2}{E_1} = \frac{(b-\frac{(a+b)*(b+d)}{N})^2}{\frac{(a+b)*(b+d)}{N}}$$

$$O_3 = c, \quad E_3 = \frac{(c+d)*(a+c)}{N} \quad \text{and} \quad \frac{(O_1-E_1)^2}{E_1} = \frac{(c-\frac{(c+d)*(a+c)}{N})^2}{\frac{(c+d)*(a+c)}{N}}$$

$$O_4 = d, \quad E_4 = \frac{(c+d)*(b+d)}{N} \quad \text{and} \quad \frac{(O_1-E_1)^2}{E_1} = \frac{(d-\frac{(c+d)*(b+d)}{N})^2}{\frac{(c+d)*(b+d)}{N}}$$

Whether or not the value of χ^2 exceeds a critical value for a preselected significance level (α) determines if the null hypothesis is rejected. This depends on the contingency table's **degrees of freedom** (df) which amounts to 1 for a 2x2 matrix. The critical χ^2 values and corresponding significance levels α for $df = 1$ are listed in Table A.5. In the experiments described in this thesis, the null hypothesis is rejected, and a difference regarded as significant, if χ^2 is greater than 3.84. This is the critical value which corresponds to the conventionally accepted significance level of 0.05 (5%). If χ^2 is greater than 3.84, then its associated probability value (p), the estimated probability of rejecting H0 when that hypothesis is in fact true, is lower than α. In this case, the alternative hypothesis is accepted. When reporting the results of each χ^2 test, both df and p values[2] are presented.

Significance level α	0.20	0.10	0.05	0.025	0.01	0.001
Critical χ^2 value	1.64	2.71	3.84	5.02	6.64	10.83

Table A.5: Critical χ^2 values and associated significance levels α for $df = 1$.

[2]Note that the lower case p stands for p value whereas the upper case P stands for precision.

Appendix B

Guidelines for Annotation of English Inclusions

B.1 Introduction

B.1.1 Ph.D. Project

The work for this Ph.D. project commenced in October 2003 and is part of the SEER project funded by the Edinburgh-Stanford Link. The overall aim of this work is to analyse the use of English inclusions in other languages, develop a classifier that can detect such inclusions in text automatically and apply the output of the recogniser to improve natural language processing (NLP) applications.

A substantial part of this work involves data annotation. Annotated data is required in order to evaluate the automatic English inclusion classifier. Moreover, double annotation is also vital for determining how feasible it is for humans to recognise English inclusions. The inter-annotator agreement which is calculated for the double annotated data serves as an upper baseline to compare the system against.

B.1.2 Annotation Guidelines

The annotation guidelines presented in this document describe the instructions and for marking up English inclusions in German text to the annotators.

Domain	Amount of data (in tokens)	% of double annotated data
Internet	32,138	100%
Space	32,237	100%
EU	32,324	100%
Total	96,699	100%

Table B.1: Amount of annotated German text per domain in tokens.

B.1.3 Annotated Data

The German data that was annotated is made up of a random selection of online newspaper articles published by Frankfurter Allgemeine Zeitung (FAZ) between 2003 and 2005. The articles stem from three distinct domains, namely internet and telecoms, space travel and European Union related subjects. Table 1 lists the number of tokens annotated for each domain and for all three domains in total as well as the portion of data that was randomly selected for double annotation. The latter will be used to determine inter-annotator agreement and therefore indicate the feasibility of the task of recognising English inclusions.

B.2 Annotation Instructions

B.2.1 General Instructions

The annotated data is used to evaluate a classifier designed to automatically identify full-word English inclusions in other languages. Therefore, this language mixing information must be recorded by the annotation.

The human annotated data serves as a gold standard to evaluate the output of the system performance. Therefore, it is crucial that the annotators' mark-up is consistent. The second purpose of this annotation, specifically the double annotation, is to determine the realistic feasibility of recognising English inclusions. This means that the human performance, measured in terms of inter-annotator agreement between two individuals, presents an upper bound for the system. As a result it is of particular importance for the annotators to adhere strictly to the guidelines.

This also means that annotators are asked not to annotate cases which they con-

sider appropriate for annotation but which are not specified by the guidelines. In such cases and should the annotator disagree with a given guideline, nevertheless follow the guidelines strictly and note down any comments, suggestions or criticisms for a specific annotation or in general for later discussion.

Annotators should perform their work independently for the purpose of measuring the real difficulty of the task. External resources like dictionaries or the web can be consulted. But in the case of further uncertainty about a specific annotation, please, take a note but follow the guidelines and work independently until the annotation process is complete. There will be opportunity to discuss difficult cases and reconcile differences after the inter-annotator agreement is determined.

B.2.2 Specific Instructions

Annotation should be done on the token level. The following types of English inclusions must be annotated: English expressions, quotes, titles, names of organisations, companies, slogans, brand names, events, products etc. as well as English abbreviations. In the following, an example is given for each category.

B.2.2.1 English Expressions

The annotators are required to mark up full-word English expressions of all grammatical categories. Annotators are also required to mark up English word forms that are part of mixed lingual hyphenated compounds as presented in Example 1. The English nouns Internet and Boom which form part of the hyphenated compound as well as the term E-Recruiting should be annotated.

(1) **E-Recruiting** ist ein Schlagwort im **Internet-Boom**-Zeitalter.

English unhyphenated compounds should also be annotated if all parts are of English origin as illustrated in Example 2.

(2)

Der Schnäppchenjäger heisst neuerdings **Smartshopper**.

B.2.2.2 English Quotes

English quotes and sayings must be annotated. See examples given below.

(3)

Think global, was ist los in der Welt?
... **God save the Queen**
... dass mit der Polizei immer noch "**Law and Order**"-Vorstellungen verbunden werden, die auch in der rechten Szene vorherrschten.

B.2.2.3 English Titles and Names

English titles, names of organisations, companies, slogans, brand names, events and products etc. must be annotated. Titles include titles of books, newspapers and films etc. but also titles of persons (see Example 4). Names include any English names given to organisations and other structures but also to things like products, satellites, events, services, slogans etc. This list is not limited as more and more English names are occurring for a variety of things. Note, however, **English geographic place names** are only to be annotated if they have a generally preferred equivalent in German. Furthermore, **English person and English-like names** are **not** to be annotated.

(4)

"Und dann erhob sich ein goldenes Wunder am Horizont" , schrieb Rudyard Kipling vor gut hundert Jahren (1889) in seinen "**Letters from the East**" ...
Der britische Rocksänger und Gitarrist von Prinz Charles zum "**Officer of the Order of the British Empire**" ernannt worden.
Prominentes Beispiel hierfür ist die Gesellschaft **Olympic Catering**.

B.2.2.4 Abbreviations

All abbreviations that expand to English definitions must be annotated if they appear together with their definition as in Example 5 or on their own.

(5)

Zeitgleich mit dem Wirtschaftsgipfel der sieben führenden Industrienationen hat am Montag in München auch der Gegengipfel "**The other economic summit**" (**TOES**) begonnen.

B.2.2.5 Pseudo-Anglicisms

Although linguists disagree on whether pseudo-anglicisms can be classed as borrowings, it is clear that such instances would not exist in the receiving language had they not been derived from the lexical item in the donor language. Therefore, they should be annotated.

(6)

Beamer, **Handy**, **Oldtimer**

B.2.2.6 Exclusions

Do **not** annotated English morphemes occurring as part of URLs, mixed-lingual unhyphenated compounds, with German inflections, or English-like person and geographic place names. Annotators are also requested not to annotate loan translations and loan words stemming from languages other than English.

URLs

URLs should not be annotated (see Example 6). English company names which appear within a URL are not to be annotated, unlike when they appear on their own.

(7)

www.ebay.de
www.stepstone.de

Mixed-lingual Unhyphenated Compounds

Mixed-lingual unhyphenated compounds should not be annotated (see Example 7).

(8)

...einen Shuttleflug ins ungewisse.

English Morphemes with German Inflections

English morphemes with German inflections, as illustrated in Example 8, should not be annotated. If the inclusion occurs without the inflection it should be annotated.

(9)

...die direkt mit den Receivern verbunden werden.

Geographic Place and Person Names

Geographic place names and person names should not be annotated (see Example 9). Note: Geographic place names are only to be annotated if their German equivalent is generally preferred in usage.

(10)

...den Reiseführer von New York.
Nach sechs Jahren George Bush ist es Zeit, ...

Loan Translations

New expressions that have entered German but that are clearly derived by translating an expression from English (or other languages) should not be annotated.

(11)

Sinn machen (English origin: make sense)
Mausklick (English origin: mouse click

Loan Words from other Languages

Loan words stemming from languages other than English should not be annotated. Although relatively rare, French, Italian, Japanese, Latin and other types of loan words do occur in German language.

(12)

...ist der unangefochtene Großmeister des Sudoku.
Wenn die Pizza keine Pizza ist.

Appendix C

TIGER Tags and Labels

C.1 Part of Speech Tag Set in TIGER

The basis for the tag set used in the TIGER annotation is the STTS tag set (Schiller *et al.*, 1995). Some minor changes to this tag set are described by Smith (2003).

ADJA	adjective, attributive
ADJD	adjective, adverbial or predicative
ADV	adverb
APPR	preposition; circumposition left
APPRART	preposition with article
APPO	postposition
APZR	circumposition right
ART	definite or indefinite article
CARD	cardinal number
FM	foreign language material
ITJ	interjection
KOUI	subordinate conjunction with *zu* and infinitive
KOUS	subordinate conjunction with sentence
KON	coordinate conjunction
KOKOM	comparative conjunction
NN	common noun
NE	proper noun

PDS	substituting demonstrative pronoun
PDAT	attributive demonstrative pronoun
PIS	substituting indefinite pronoun
PIAT	attributive indefinite pronoun with/without determiner
PPER	non-reflexive personal pronoun
PPOSS	substituting possessive pronoun
PPOSAT	attributive possessive pronoun
PRELS	substituting relative pronoun
PRELAT	attributive relative pronoun
PRF	reflexive personal pronoun
PWS	substituting interrogative pronoun
PWAT	attribute interrogative pronoun
PWAV	adverbial interrogative or relative pronoun
PROAV	pronominal adverb
PTKZU	*zu* before infinitive
PTKNEG	negative particle
PTKVZ	separable verbal particle
PTKANT	answer particle
PTKA	particle with adjective or adverb
SGML	SGML markup
SPELL	letter sequence
TRUNC	word remnant
VVFIN	finite verb, full
VVIMP	imperative, full
VVINF	infinitive, full
VVIZU	infinitive with *zu*, full
VVPP	perfect participle, full
VAFIN	finite verb, auxiliary
VAIMP	imperative, auxiliary
VAINF	infinitive, auxiliary
VAPP	perfect participle, auxiliary
VMFIN	finite verb, modal
VMINF	infinite verb, modal

VMPP	perfect participle, modal
XY	non-word containing non-letter
$,	comma
$.	sentence-final punctuation mark
$(	other sentence-internal punctuation mark

C.2 Phrase Category (Node) Labels in TIGER

AA	superlative phrase with *am*
AP	adjective phrase
AVP	adverbial phrase
CAC	coordinated adposition
CAP	coordinated adjective phrase
CAVP	coordinated adverbial phrase
CCP	coordinated complementiser
CH	chunk
CNP	coordinated noun phrase
CO	coordination
CPP	coordinated adpositional phrase
CS	coordinated sentence
CVP	coordinated verb phrase (non-finite)
CVZ	coordinated infinitive with *zu*
DL	discourse level constituent
ISU	idiosyncratic unit
MTA	multi-token adjective
NM	multi-token number
NP	noun phrase
PN	proper noun
PP	adpositional phrase
QL	quasi-language
S	sentence
VP	verb phrase (non-finite)
VZ	infinitive with *zu*

C.3 Grammatical Function (Edge) Labels in TIGER

AC	adpositional case marker
ADC	adjective component
AG	genitive attribute
AMS	measure argument of adjective
APP	apposition
AVC	adverbial phrase component
CC	comparative complement
CD	coordinating conjunction
CJ	conjunct
CM	comparative conjunction
CP	complementiser
CVC	collocational verb construction
DA	dative
DH	discourse-level head
DM	discourse marker
EP	expletive *es*
HD	head
JU	junctor
MNR	postnominal modifier
MO	modifier
NG	negation
NK	noun kernel element
NMC	numerical component
OA	accusative object
OA	second accusative object
OC	clausal object
OG	genitive object
OP	prepositional object
PAR	parenthetical element

PD	predicate
PG	phrasal genitive
PH	placeholder
PM	morphological particle
PNC	proper noun component
RC	relative clause
RE	repeated element
RS	reported speech
SB	subject

Bibliography

Abresch, J. (2007). *Englisches in Gesprochenem Deutsch. Eine empirische Analyse der Aussprache und Beurteilung englischer Laute im Deutschen.* Universität Bonn, Bonn. Ph.D. Thesis.

Agirre, E. and 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*, pages 11–19, Luxembourg.

Ahmed, B., Cha, S.-H., and Tappert, C. (2004). Language identification from text. Using n-gram based cumulative frequency addition. In *Proceedings of CSIS Research Day*, Pace University, New York, USA.

Ahn, K., Alex, B., Bos, J., Dalmas, T., Leidner, J. L., and Smillie, M. B. (2004). Cross-lingual question answering with QED. In *Workshop of the Cross-Lingual Evaluation Forum (CLEF) at the European Conference for Digital Libraries (ECDL-2004)*, pages 335–342, Bath, UK.

Alex, B. (2005). An unsupervised system for identifying English inclusions in German text. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), Student Research Workshop*, pages 133–138, Ann Arbor, Michigan, USA.

Alex, B. (2006). Integrating language knowledge resources to extend the English inclusion classifier to a new language. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 595–600, Genoa, Italy.

Alex, B. and Grover, C. (2004). An XML-based tool for tracking English inclusions in German text. In *Proceedings of PAPILLON 2004 - Workshop on Multilingual Lexical Databases*, Grenoble, France.

Alex, B., Dubey, A., and Keller, F. (2007). Using foreign inclusion detection to improve parsing performance. In *Proceedings of EMNLP-CoNLL 2007*, pages 151–160, Prague, Czech Republic.

Allen, J., Hunnicutt, M. S., and Klatt, D. H. (1987). *From text to speech: the MITalk system.* Cambridge University Press, Cambridge, UK.

Andersen, G. (2005). Assessing algorithms for automatic extraction of anglicisms in Norwegian texts. In *Proceedings of the International Conference on Corpus Linguistics (CL2005)*, Birmingham, UK. Available online at: `http://www.corpus.bham.ac.uk/PCLC/`.

Androutsopoulos, J. K., Bozkurt, N., Breninck, S., Kreyer, C., Tronow, M., and Tschann, V. (2004). Sprachwahl im Werbeslogan. Zeitliche Entwicklung und branchenspezifische Verteilung englischer Slogans in der Datenbank von slogans.de. *NET.WORX*, **41**, 4–27. Available online at: `http://www.mediensprache.net/de/networx/docs/networx-41.asp`.

Badino, L., Barolo, C., and Quazza, S. (2004). A general approach to TTS reading of mixed-language texts. In *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004 - ICSLP)*, pages 849–852, Jeju Island, Korea.

Bae, H. S. and L'Homme, M.-C. (2006). Converting a monolingual lexical database into a multilingual specialized dictionary. In *Proceedings of the Conference on Multilingualism and Applied Comparative Linguistics*, Brussels, Belgium.

Baldauf, S. (2004). A Hindi-English jumble. *The Christian Science Monitor*. Published on 23/11/2004. Available online at: `http://www.csmonitor.com/2004/1123/p01s03-wosc.html`.

Bartsch, S. and Siegrist, L. (2002). Anglizismen in Frachsprachen des Deutschen. Eine Untersuchung auf Basis des Darmstädter Corpus Deutscher Fachsprachen. *Muttersprache*, **112**(4), 309–323.

BBC News (2006). Beijing stamps out poor English. In *BBC News, Asia-Pacific*. 15th of October 2006. Available online at: `http://news.bbc.co.uk/1/hi/world/asia-pacific/6052800.stm`.

Beesley, K. R. (1988). Language Identifier: A computer program for automatic natural-language identification of on-line text. In *Proceedings of the 29th Annual Conference of the American Translators Association*, pages 47–54, Medford, New Jersey, USA.

Berns, M. (1992). Bilingualism with English as the other tongue: English in the German legal domain. *World Englishes*, **11**(2/3), 155–161.

Betz, W. (1936). *Der Einfluß des Lateinischen auf den althochdeutschen Sprachschatz*. Winter, Heidelberg.

Betz, W. (1974). Lehnwörter und Lehnprägungen im Vor- und Frühdeutschen. In F. Maurer and H. Rupp, editors, *Deutsche Wortgeschichte*, volume 1, pages 135–163. Walter de Gruyter, Berlin/New York.

Björkman, M. and Gösta, E. (1957). *Experimentalpsykologiska metoder*. Almqvist & Wiksell/Gebers förlag AB, Stockholm.

Black, A. W. and Taylor, P. A. (1997). The Festival speech synthesis system: System documentation. Technical report, Human Communciation Research Centre, University of Edinburgh, Scotland, UK.

Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., J. Klavans, M. L., Marcus, M., Roukos, S., Santorini, B., and Strzalkowski., T. (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. In E. Black, editor, *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 306–311, Pacific Grove, California, USA.

Bohn, O.-S. and Flege, J. E. (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics*, **11**, 303–328.

Bohn, O.-S. and Flege, J. E. (1992). The production of new and similar vowels by adult German learners of english. *Studies in Second Language Acquisition*, **15**, 131–158.

Booth, T. L. and Thompson, R. A. (1973). Applying probability measures to abstract languages. *IEEE Transactions on Computers*, **C-22**(5), 442–450.

Boyd, S. (1993). Attrition or expansion? Changes in the lexicon of Finnish and American adult bilinguals in Sweden. In K. Hyltenstam and Å. Viberg, editors, *Progression and regression in language: Sociocultural, neuropsychological & linguistic perspectives*. Cambridge University Press, Cambridge.

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, September 20-21 (TLT02)*, pages 24–41, Sozopol, Bulgaria.

Brants, T. (2000a). Inter-annotator agreement for a German newspaper corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, pages 1435–1439, Athens, Greece.

Brants, T. (2000b). TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP 2000)*, pages 224–231, Seattle, Washington, USA.

Breen, J. (2005). Expanding the lexicon: Harvesting neologisms in Japanese. In *Proceedings of the Papillon (Multi-lingual Dictionary) Project Workshop*, Chiang Rai, Thailand. Available online at: `http://www.csse.monash.edu.au/~jwb/neolharv.html`.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Blemont, California.

Bresnan, J. (2001). *Lexical-Functional Syntax.* Vol. 16 of Textbooks in Linguistics. Blackwell, Oxford.

Brown, N. (1996). *Russian Learner's Dictionary: 10,000 Russian Words in Order of Frequency.* Routledge, New York, 1st edition.

Busse, U. (1993). *Anglizismen im Duden - Eine Untersuchung zur Darstellung englischen Wortguts in den Ausgaben des Rechtschreibdudens von 1880-1986.* Niemeyer, Tübingen.

Busse, U. (1994). 'Wenn die Kötterin mit dem Baddibuilder...'. Ergebnisse einer Informantenbefragung zur Aussprache englischer Wörter im Deutschen. In D. Halwachs and I. Stütz, editors, *Sprache - Sprechen - Handeln*, pages 23–30. Tübingen, Niemeyer.

Busse, U. and Görlach, M. (2002). German. In M. Görlach, editor, *English in Europe*, pages 13–36. Oxford University Press, Oxford.

Butt, M., Dyvik, H., King, T. H., Masuichi, H., and Rohrer, C. (2002). The Parallel Grammar Project. In *Proceedings of the COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7, Taipee, Taiwan.

Callahan, L. (2004). *Spanish/English Codeswitching in a Written Corpus.* John Benjamins, Amsterdam/Philadelphia.

Campbell, N. (1998). Foreign-language speech synthesis. In *Proceedings of 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia.

Capstik, J., Diagne, A. K., Erbach, G., Uszkoreit, H., Leisenberg, A., and Leisenberg, M. (1999). A system for supporting cross-lingual information retrieval. *Information Processing and Management - Special topic issue, Web Research and Information Retrieval*, **36**(2), 275–289.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, **22**(2), 249–254.

Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., and Voormann, H. (2003). The NITE XML toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, **35**(3), 353–363.

Carstensen, B. (1963). Amerikanische Einflüsse auf die deutsche Sprache. *Jahrbuch für Amerikastudien*, **8**, 35–55. Reprinted in: Carstensen, B. and Galinsky, H. (1967). Amerikanismen der deutschen Gegenwartssprache: Entlehnungsvorgänge und ihre stilistischen Aspekte. Winter, Heidelberg.

Carstensen, B. (1965). Englische Einflüsse auf die deutsche Sprache nach 1945. *Beihefte zum Jahrbuch für Amerikastudien 13.* Winter, Heidelberg.

Carstensen, B. (1979). Evidente und latente Einflüsse des Englischen auf das Deutsche. In P. Braun, editor, *Fremdwort-Diskussion*, pages 90–94. Fink, Munich.

Carstensen, B. (1986). Euro-English. In D. Kastovsky, editor, *Contrastive and Applied Linguistics*, pages 827–835. De Gruyter, Berlin.

Carstensen, B., Griesel, H., and Meyer, H.-G. (1972). Zur Intensität des englischen Einflusses auf die deutsche Pressesprache. *Muttersprache*, **82**(4), 238–243.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94)*, pages 161–175, Las Vegas, Nevada.

CCITT (1989). Telephone transmission quality. *Blue Book, Series P Recommendations*, **5**.

Celex (1993). *The CELEX lexical database*. Centre for Lexical Information, Max Planck Institute of Psycholinguistics. Available online at: `http://www.kun.nl/celex/`.

Chang, Y. (2005). *Anglizismen in der deutschen Fachsprache der Computertechnik*. Peter Lang, Frankfurt am Main.

Clyne, M. (1993). *Dynamics of Language Contact: English and Immigrant Languages*. Cambridge University Press, Cambridge.

Clyne, M. G. (1997). *The German language in a changing Europe*. Cambridge University Press, Cambridge.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A., and Zue, V., editors (1997). *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge.

Corr, R. (2003). *Anglicisms in German Computing Technology*. Trinity College Dublin, Computer Science Department, Dublin, B.A. Dissertation.

Crystal, D. (2001). *Language and the Internet*. Cambridge University Press, Cambridge.

Dalrymple, M. (2001). *Lexical Functional Grammar*. Vol. 34 of Syntax and Semantics. Academic Press, New York.

Dalrymple, M., Kaplan, R., Maxwell, J., and Zaenen, A., editors (1995). *Formal Issues in Lexical-Functional Grammar*. No. 47 in CSLI Lecture Notes. CSLI, Stanford, California.

Damashek, M. (1995). Gauging similarity with n-grams: Language independent categorization of text. *Science*, **267**(5199), 843–848.

Dipper, S. (2003). *Implementing and Documenting Large-scale Grammars - German LFG*. Ph.D. thesis, IMS, University of Stuttgart, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS).

Dubey, A. (2005a). *Statistical Parsing for German: Modeling syntactic properties and annotation differences*. Ph.D. thesis, Saarland University, Germany.

Dubey, A. (2005b). What to do when lexicalization fails: parsing German with suffix analysis and smoothing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 314–321, Ann Arbor, Michigan, USA.

Dubey, A. and Keller, F. (2003). Parsing German with sister-head dependencies. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 96–103, Sapporo, Japan.

Duckworth, D. (1977). Zur terminologischen und systematischen Grundlage der Forschung auf dem Gebiet der englisch-deutschen Interferenz. In H. Kolb and H. Lauffer, editors, *Sprachliche Interferenz*, pages 36–56. Niemeyer, Tübingen.

Dumont, E. (2005). Anti-spam: Microsoft choisit de passer en force pour imposer Sender ID. *ZDNet France*. Available online at: `http://zdnet.fr/actualites/internet/0,39020774,39235930,00.htm`.

Dunger, H. (1882). *Wörterbuch von Verdeutschungen entbehrlicher Fremdwörter mit besonderer Berücksichtigung der von dem Großen Generalstabe, im Postwesen und in der Reichsgesetzgebung angenommenen Verdeutschungen*. Teubner, Leipzig. Reprint published in 1989, W. Viereck, editor, Olms, Hildesheim.

Dunger, H. (1909). *Wider die Engländerei in der deutschen Sprache*. Allgemeiner Deutscher Sprachverein, Berlin. Reprint published in 1989, W. Viereck, editor, Olms, Hildesheim.

Dunn, J. (2007). Face control, electronic soap and the four-storey cottage with a jacuzzi: anglicisation, globalisation and the creation of linguistic difference. In R. Fischer and H. Pulaczewska, editors, *Proceedings of the international conference Anglicisms in Europe*. Cambridge Scholars Publishing, Newcastle-upon-Tyne. Forthcoming.

Dunning, T. (1994). Statistical identification of language. Technical report, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico, USA.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht.

Dutoit, T. and Leich, H. (1993). Text-to-speech synthesis based on a MBE re-synthesis of the segments database. *Speech Communication*, **13**, 435–440.

Eckert, T., Johann, A., Känzig, A., Künig, M., Müller, B., Schwald, C., and Walder, L. (2004). Is English a 'killer language'? The globalisation of a code. *eHistLing*, **1**. Available online at: `http://www.ehistling.meotod.de/data/papers/group_g_pub.pdf`.

Ejerhed, E., Källgren, G., Wennstedt, O., and Åström, M. (1992). The linguistic annotation of the Stockholm-Umeå corpus project. Technical report, Department of General Linguistics, University of Umeå, Umeå, Sweden.

Eklund, R. and Lindström, A. (1996). Pronunciation in an internationalized society: a multi-dimensional problem considered. In *Proceedings of Fonetik 1996*, pages 123–126, Nässlingen, Sweden.

Eklund, R. and Lindström, A. (1998). How to handle "foreign" sounds in Swedish text-to-speech conversion: Approaching the 'xenophone' problem. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 1998)*, pages 2831–2834, Sydney, Australia.

Eklund, R. and Lindström, A. (2001). Xenophones: An investigation of phone set expansion in swedish and implications for speech recognition and speech synthesis. *Speech Communication*, **35**(1-2), 81–102.

Etiemble, R. (1964). *Parlez-vous franglais?* Gallimard, Paris.

Falk, Y. N. (2001). *Lexical-Functional Grammar, An Introduction to Parallel Constraint-Based Syntax*. No. 126 in CSLI Lecture Notes. CSLI, Stanford, California.

Farrugia, P.-J. (2005). *Text to Speech Technologies for Mobile Telephony Services*. Department of Computer Science and AI, University of Malta, Msida, Malta. MSc Thesis.

FAZ-Magazin (1996). Jil Sander spricht, interview with Jil Sander. *FAZ-Magazin*, page 21. Published on 12/03/1996.

Fink, H. (1980). Zur Aussprache von Angloamerikanischem im Deutschen. In W. Viereck, editor, *Studien zum Einfluß der englischen Sprache auf das Deutsche*, pages 109–183. Gunter Narr, Tübingen.

Fink, H., Fijas, L., and Schons, D. (1997). *Anglizismen in der Sprache der Neuen Bundesländer: eine Analyse zur Verwendung und Rezeption*. Peter Lang, Frankfurt am Main.

Finkel, J., Dingare, S., Manning, C. D., Nissim, M., Alex, B., and Grover, C. (2005). Exploring the boundaries: Gene and protein identification in biomedical text. *BMC Bioinformatics*, **6**(Suppl1), S5.

Fischer, R. and Pulaczewska, H., editors (2007). *Proceedings of the international conference Anglicisms in Europe*. Cambridge Scholars Publishing, Newcastle-upon-Tyne. Forthcoming.

Fitt, S. (1998). *Processing Unfamiliar Words: A Study in the Preception and Production of Native and Foreign Placenames*. University of Edinburgh, Edinburgh. Ph.D. Thesis.

Flaitz, J. (1993). French attitudes toward the ideology of English as an international language. *World Englishes*, **12**(2), 179–191.

Flege, J. E. (1987). Effects of equivalence classification on the production of foreign language speech sounds. In J. A. and L. J., editors, *Sound Patterns in Second Language Acquisition*. Foris, Dordrecht.

Flege, J. E. (1997). English vowel production by Dutch talkers: more evidence for the "similar" vs. "new" distinction. In A. James and J. Leather, editors, *Second-Language Speech: Structure and Process*, pages 11–52. Mouton de Gruyter, Berlin/New York.

Flege, J. E., Bohn, O.-S., and Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, **25**, 437–470.

Font-Llitjós, A. and Black, A. W. (2001). Knowledge of language origin improves pronunciation accuracy of proper names. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 1919–1922, Aalborg, Denmark.

Forst, M. and Kaplan, R. M. (2006). The importance of precise tokenizing for deep grammars. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 369–372, Genoa, Italy.

Furiassi, C. (2007). Non-adapted anglicisms in Italian: Attitudes, frequency counts, and lexicographic implications. In R. Fischer and H. Pulaczewska, editors, *Proceedings of the international conference Anglicisms in Europe*. Cambridge Scholars Publishing, Newcastle-upon-Tyne. Forthcoming.

Furiassi, C. and Hofland, K. (2007). *The Retrieval of False Anglicisms in Newspaper Texts*. Rodopi, Amsterdam/New York.

Galinsky, H. (1980). American neologisms in German. *American Speech*, **55**(4), 243–263.

Gardner-Chloros, P. (1995). Code-switching in community, regional and national repertories: The myth of the discreteness of linguistic systems. In L. Milroy and P. Muysken, editors, *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, pages 68–89. Cambridge University Press, Cambridge.

Gentsch, K. (1994). *English borrowings in German newspaper language: Motivations, frequencies, and types, on the basis of the Frankfurter Allgemeine Zeitung, Münchner Merkur, and Bild*. Dissertation, Swarthmore College, Pennsylvania, USA.

Glahn, R. (2000). *Der Einfluß des Englischen auf gesprochene deutsche Gegenwartssprache*. Peter Lang, Frankfurt am Main.

Goldstein, M. (1995). Listeners are required to indicate their preferred presentation from each pair. *Speech Communication*, **16**(3), 225–244.

Görlach, M. (1994). Continental pun-dits. *English Today*, **37, 10**(1), 50–52.

Görlach, M. (2001). *Dictionary of European Anglicisms: A Usage Dictionary of Anglicims in Sixteen European Languages*. Oxford University Press, Oxford.

Graddol, D. (1997). *The future of English?* British Council, London.

Grefenstette, G. (1995). Comparing two language identification schemes. In *Proceedings of the 3rd International Conference on Statistical Analysis of Textual Data (JADT 1995)*, pages 263–268, Rome, Italy.

Grefenstette, G. (1999). The WWW as a resource for example-based machine translation tasks. In *Proceedings of ASLIB'99 Translating and the Computer*, London, UK.

Grefenstette, G. and Nioche, J. (2000). Estimation of English and non-English language use on the WWW. In *Proceedings of RIAO (Recherche d'Informations Assistée par Ordinateur) 2000*, pages 237–246, Paris, France.

Greisbach, R. (2003). The pronunciation of loanwords in German. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, pages 799–802, Barcelona, Spain.

Grover, C., Matheson, C., Mikheev, A., and Moens, M. (2000). LT TTT - a flexible tokenisation tool. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 1147–1154, Athens, Greece.

Grover, C., Matthews, M., and Tobin, R. (2006). Tools to address the interdependence between tokenisation and standoff annotation. In *Proceedings of NLPXML 2006*, pages 19–26, Trento, Italy.

Guiraud, P. (1965). *Les mots étrangers*. Presses universitaires de France, Paris.

Hachey, B., Alex, B., and Becker, M. (2005). Investigating the effects of selective sampling on the annotation task. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*, pages 144–151, Ann Arbor, Michigan, USA.

Hall-Lew, L. A. (2002). *English Loanwords in Mandarin Chinese.* University of Arizona, Tucson, US. BA Honors Thesis.

Harris, G. (2003). Global English and German today. Occasional paper. Centre for Language in Education, University of Southhampton, Southampton, UK.

Haugen, E. (1950). The analysis of linguistic borrowing. *Language*, **26**, 210–231.

Hedderich, N. (2003). Language Change in Business German. *Global Business Languages*, **8**, 47–55. C. E. Keck and A. G. Wood, editors.

Henrich, P. (1988). Ansatz zur automatischen Graphem-zu-Phonem-Umsetzung von Fremdwörtern in einem Deutschsprachigen Vorleseautomaten für unbegrenzten Wortschatz. In *Fortschritte der Akustik - DAGA '88*, pages 661–664, Bad Honnef, Germany.

Hilgendorf, S. K. (1996). The impact of English in Germany. *English Today*, **12**(3), 3–14.

Hilgendorf, S. K. (2007). English in Germany: contact, spread and attitudes. *World Englishes*, **26**(2), 131–148.

Hoberg, R. (2000). Sprechen wir bald alle Denglisch oder Germeng? In K. M. Eichhoff-Cyrus and R. Hober, editors, *Die deutsche Sprache zur Jahrtausendwende.* Dudenverlag, Mannheim.

Hohenhausen, P. (2001). Neuanglodeutsch. Zur vermeintlichen Bedrohung des Deutschen durch das Englische. *German as a foreign language*, **1**, 57–87.

Holmes, J. and Holmes, W. (2001). *Speech Synthesis and Recognition.* Taylor & Francis, New York, 2nd edition.

Huffman, J. P. (1998). *Family, commerce and religion in London and Cologne : Anglo-German emigrants, c.1000-c.1300.* Cambridge University Press, Cambridge.

Humbley, J. (2006). Anglizismen im Französischen: immer noch ein Sonderfall? In *Proceedings of the international conference Anglicisms in Europe*, Universität Regensburg, Germany.

Internet World Stats (2007). Internet top 10 languages. In *Internet Worlds Stats.* Published on 19/03/2007. Available online at: `http://www.internetworldstats.com/stats7.htm`.

Jabłoński, M. (1990). *Regularität und Variabilität in der Rezeption englischer Internationalismen im modernen Deutsch, Französisch und Polnisch : aufgezeigt in den Bereichen Sport, Musik und Mode*. M. Niemeyer, Tübingen.

Jacquemin, C. and 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.

John T. Maxwell, I. and Kaplan, R. (1993). The interface between phrasal and functional constraints. *Computational Linguistics*, **19**, 571–589.

Johnson, S. (1993). Solving the problem of language recognition. Technical report, School of Computer Studies, University of Leeds, Leeds, UK.

Joshi, A. K. (1982). Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th International Conference on Computational Linguistics (COLING 1982)*, pages 145–150, Prague, Czechoslovakia.

Junker, G. H. (2006). *Der Anglizismen-Index. Anglizismen, Gewinn oder Zumutung?* Ifb, Paderborn.

Kachru, Y. (2006). Mixers lyricing in Hinglish: blending and fusion in Indian pop culture. *World Englishes*, **25**(2), 223–233.

Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, **29**(3), 458–484.

Kikui, G. (1996). Identifying the coding system and language of on-line documents on the internet. In *Proceedings of International Conference On Computational Linguisitics (COLING 1996)*, page 652657, Copenhagen, Denmark.

Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, **10**(2), 135–155.

Kilgarriff, A. (2003). What computers can and cannot do for lexicography, or us precision, them recall. Technical Report ITRI-03-16, Information Technology Research Institute, University of Brighton. Also published in *Proceedings of the 3rd Conference of the Asian Association for Lexicography (ASIALEX 2003)*, Tokyo, Japan.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on web as corpus. *Computational Linguistics*, **29**(3).

Kilgarriff, A., Rundell, M., and Dhonnachdha, U. (2005). Corpus creation for lexicography. In *Proceedings of ASIALEX 2005*, Singapore.

King, R. E. (2000). *The lexical basis of grammatical borrowing: a Prince Edward Island French case study*. John Benjamins, Amsterdam/Philadelphia.

Kirkness, A. (1984). Aliens, denziens, hybrids and natives: foreign influence on the etymological structure of German vocabulary. In C. V. J. Russ, editor, *Foreign Influences on German*. Lochee, Dundee.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 423–430, Saporo, Japan.

Klein, D., Smarr, J., Nguyen, H., and Manning, C. D. (2003). Named entity recognition with character-level models. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL 2003)*, Edmonton, Canada.

Koehn, P. (2004). Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings the 6th Conference of the Association for Machine Translation in the Americas (AMTA 2004)*, pages 115–124, Washington DC, USA.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of EACL 2003*, pages 187–194, Budapest, Hungary.

Koekkoek, B. J. (1958). A note on the German borrowing of American brand names. *American Speech*, **33**(3), 236–237.

Kranig, S. (2005). *Evaluation of Language Identification Methods*. Universität Tübingen, Tübingen, Germany. BA Honors Thesis.

Kupper, S. (2003). *Anglizismen in deutschen und französischen Werbeanzeigen. Zum Umgang von Deutschen und Franzosen mit Anglizismen*. Tectum Verlag, Marburg.

Landis, R. J. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.

Laroch-Claire, Y. (2004). *Évitez le franglais, parlez français*. Albin Michel, Paris.

Larson, M., Willett, D., Köhler, J., and Rigoll, G. (2000). Compound splitting and lexical recombination for improved performance of a speech recognition system for German parliamentary speeches. In *Proceeding of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, pages 945–948, Beijng, China.

Latorre, J., Iwano, K., and Furui, S. (2005). Polyglot synthesis using a mixture of monolingual corpora. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, pages 1–4, Philadelphia, Pennsylvania, USA.

Lewis, S., McGrath, K., and Reuppel, J. (2004). Language identification and language specific letter-to-sound rules. *Colorado Research in Linguistics*, **17**(1).

Lin, D. (1995). A dependency-based method for evaluating broad-coverage parsers. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 1420–1425, Montreal, Canada.

Lindström, A. and Eklund, R. (1999a). [jɑ̀ːmes] or [dʒɛɪmz] or perhaps something in-between? Recapping three years of xenophone studies. In *Proceedings of Fonetik 1999*, pages 109–112, University of Gothenburg, Sweden.

Lindström, A. and Eklund, R. (1999b). Xenophones revisited: Linguistic and other underlying factors affecting the pronunciation of foreign items in Swedish. In *Proceedings of the International Conference of Phonetic Sciences (ICPhS 1999)*, pages 2227–2230, Berkeley, California, USA.

Lindström, A. and Eklund, R. (2000). How foreign are "foreign" speech sounds? implications for speech recognition and speech synthesis. In *Muli-Lingual Interoperability in Speech Technology, RTO Meeting Proceedings 28*, Hull (Québec), Canada.

Lindström, A. and Eklund, R. (2002). Xenophenomena: studies of foreign language influence at several linguistic levels. In *Proceedings of the 24. Jahrestag der Deutschen Gesellschaft für Sprachwissenschaft: Mehrsprachigkeit Heute. AG 8: Integration Fremder Wörter*, pages 132–134, Universität Mannheim, Mannheim, Germany.

Lindström, A. and Kasaty, A. (2000). A two-level approach to the handling of foreign items in Swedish speech technology applications. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, volume 1, pages 54–57, Beijing, China.

Link, R. (2004). Zwischen Denglisch und Altfränkisch. In *Hintergrund Politik*. Deutschlandfunk. Broadcasted on 24/04/2004.

Ljung, M. (1998). *Sinkheads, hacjers och lams ankor*. Bokförlaget Trevi, Stockholm.

Lohmann, K., Uphoff, M., and Kattner, C. (2004). European anglicisms. Katholishe Universität Eichstätt-Ingolstadt, Germany. Available online at: `http://66.102.9.104/search?q=cache:u7w-nUk-mpgJ:www1.ku-eichstaett.de/S%LF/EngluVglSW/schule30.pdf+Lohmann+%22European+Anglicisms%22&hl=en&ct=clnk&cd=%1&gl=uk&client=firefox-a`.

Mamère, N. and Warin, O. (1999). *Non merci, Oncle Same!* Ramsay, Paris.

Manning, C. D. and Schütze, H. (2001). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

Marcadet, J. C., Fischer, V., and Waast-Richard, C. (2005). A transformation-based learning approach to language identification for mixed-lingual text-to-speech synthesis. In *Proceedings of the 9th European Conference on Speech Communication*

and Technology (Interspeech 2005 - Eurospeech), pages 2249–2252, Lisbon, Portugal.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

Markel, J. D. and H., G. A. (1976). *Linear Prediction of Speech*. Springer Verlag, Berlin/New York.

McClure, E. and McClure, M. (1988). Macro- and micro-sociolinguistic dimensions of code-switching in Vingard. In M. Heller, editor, *Codeswitching: Anthropological and Sociolinguistic Perspectives*, pages 25–51. Mouton de Gruyter, Berlin/New York.

Mihalcea, R. and Moldovan, D. I. (1999). A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, pages 152–158, Maryland, New York, USA.

Modjeska, N., Markert, K., and Nissim, M. (2003). Using the web in machine learning for other-anaphora resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, Sapporo, Japan.

Monz, C. and de Rijke, M. (2001). Shallow morphological analysis in monolingual information retrieval for Dutch, German, and Italian. In *Second Workshop of the Cross-Language Evaluation Forum (CLEF 2001)*, pages 262–277, Darmstadt, Germany.

Moody, A. J. (2006). English in Japanese popular culture and J-Pop music. *World Englishes*, **25**(2), 209–222.

Moser, H. (1985). Die Entwicklung der deutschen Sprache seit 1945. In W. Besch, O. Reichmann, and S. Sonderegger, editors, *Sprachgeschichte: ein Handburch zur Geschichte der deutschen Sprache und ihrer Erforschung*, pages 1678–1707. De Gruyter, Berlin.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, **9**(5/6), 453–467.

Müller, C. (2003). Anglizimen in verschiedenen EU-Sprachen. Katholishe Universität Eichstätt-Ingolstadt, Germany. Available online at: `http://www1.ku-eichstaett.de/SLF/EngluVglSW/schule15.pdf`.

Muysken, P. (2000). *Bilingual Speech: A Typology of Code-Mixing*. Cambridge University Press, Cambridge.

Myers-Scotton, C. (1993). *Duelling Languages: Grammatical Structure in Code-Switching*. Clarendon Press, Oxford.

New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behaviour Research Methods Instruments & Computers*, **36**(3), 516–524.

Newman, P. (1987). Foreign language identification: First step in the translation process. In *Proceedings of the 28th Annual Conference of the American Translators Association*, pages 509–516, Medford, New Jersey.

Nicholls, D. (2003). False friends between French and English. *MED Magazine*, **9**. Available online at: `http://www.macmillandictionary.com/med-magazine/July2003/09-French-Engl%ish-false-friends.htm`.

Nusbaum, H. C., Schwab, E. C., and Pisoni, D. B. (1984). Subjective evaluation of synthetic speech: Measuring preference, naturalness and acceptability. Technical report, Speech Research Laboratory, Indiana University, USA.

Onysko, A. (2006). English codeswitching in the German newsmagazine Der Spiegel. In R. Muhr, editor, *Innovation and Continuity in Language and Communication of Different Language Cultures*, pages 261–290. Peter Lang, Frankfurt am Main.

Onysko, A. (2007). *Anglicisms in German: Borrowing, Lexical Productivity and Written Codeswitching*. De Gruyter, Berlin/New York.

Padró, M. and Padró, L. (2004). Comparing methods for language identification. *Procesamiento del Lenguaje Natural*, **33**, 155–162.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: a method for automatic evaluation of machine translation. Technical report, IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, New York, USA.

Pergnier, M. (1989). *Les anglicismes*. Presse Universitaires de France, Paris.

Petrakis, H. M. (1965). The talkies - handie and walkie. Available online at: `http://www.batnet.com/mfwright/hthistory.html`.

Pfister, B. and Romsdorfer, H. (2003). Mixed-lingual analysis for polyglot TTS synthesis. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 2037–2040, Geneva, Switzerland.

Picone, M. D. (1996). *Anglicisms, Neologisms and Dynamic French*. John Benjamins, Amsterdam/Philadelphia.

Plümer, N. (2000). *Anglizismus - Purismus - Sprachliche Identität: eine Untersuchung zu den Anglizismen in der deutschen und französischen Mediensprache*. Peter Lang, Frankfurt am Main.

Poplack, S. (1988). Contrasting patterns of code-switching in two communities. In M. Heller, editor, *Codeswitching: Anthropological and Sociolinguistic Perspectives*, pages 215–244. Mouton de Gruyter, Berlin/New York.

Poplack, S. (1993). Variation theory and language contact. In D. Preston, editor, *American Dialect Research*, pages 25–286. John Benjamins, Amsterdam, Philadelphia.

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

Qu, Y. and Grefenstette, G. (2004). Finding ideographic representations of Japanese names written in Latin script via language identification and corpus validation. In *Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 183–190, Barcelona, Spain.

Rabiner, L. R. (1989). A tutorial in Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), 257–286.

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In *Proceedings of the 3rd Workshop on Very Large Corpora (ACL 1995)*, pages 82–94, Cambridge, Massachusetts, USA.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, pages 527–534, Maryland, New York, USA.

Rohrer, C. and Forst, M. (2006). Improving coverage and parsing quality of a large-scale LFG for German. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2206–2211, Genoa, Italy.

Rollason, C. (2005). Language borrowings in a context of unequal systems: Anglicisms in French and Spanish. *Lingua Franca, Le Bulletin des Interprètes du Parlement Européen*, **8**(2), 9–14.

Romsdorfer, H. and Pfister, B. (2003). Mixed-lingual text analysis for polyglot TTS synthesis. Presentation for IM2 Summer Institute, Crans-Montana, Switzerland.

Romsdorfer, H. and Pfister, B. (2004). Multi-context rules for phonological processing in polyglot TTS synthesis. In *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004 - ICSLP)*, pages 845–848, Jeju Island, Korea.

Romsdorfer, H., Pfister, B., and Beutler, R. (2005). A mixed-lingual phonological component which drives the statistical prosody control of a polyglot TTS synthesis system. In S. Bengio and H. Bourlard, editors, *Machine Learning for Multimodal Interaction*, pages 263–276. Springer, Heidelberg.

Sankoff, D. and Poplack, S. (1984). Borrowing: the synchrony of integration. *Linguistics*, **22**, 99–136.

Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Draft, Universität Stuttgart and Universität Tübingen.

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

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL 1995 SIGDAT Workshop*, pages 47–50, Dublin, Ireland.

Schütte, D. (1996). *Das schöne Fremde. Anglo-amerikanische Einflüsse auf die Sprache der deutschen Zeitschriftenwerbung*. Westdeutscher Verlag, Opladen.

Schwartz, A. and Hearst, M. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the Pacific Symposium on Biocomputing (PSB 2003)*, pages 451–462, Kauai, Hawaii.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois.

Shirai, S. (1999). *Gemination in loans from English to Japanese*. University of Washington, US. MA Thesis.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 1992)*, volume 2, pages 867–870, Banff, Canada.

Sinclair, J., editor (1995). *Collins COBUILD English Language Dictionary*. HarperCollins, London, UK, 2nd edition.

Sjöbergh, J. (2003). Combining POS-taggers for improved accuracy on Swedish text. In *Proceedings of NoDaLiDa 2003*, Reykjavik, Iceland.

Skut, W., Krenn, B., Brants, T., and Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP 1997)*, Washington, DC, USA.

Skut, W., Brants, T., Krenn, B., and Uszkoreit, H. (1998). A linguistically interpreted corpus of German newspaper text. In *Proceedings of the Conference on Language Resources and Evaluation (LREC 1998)*, pages 705–712, Granada, Spain.

Smith, G. (2003). A brief introduction to the TIGER Treebank. Version 1. Technical report, Universität Potsdam, Potsdam, Germany.

Sokol, D. K. (2000). An informal look at the invasion of anglicisms and americanisms in modern French. *LINGUA NON GRATA*, (24). Available online at: `http://www.eltnewsletter.com/back/August2000/art242000.htm`.

Stein, A. and Schmidt, H. (1995). Étiquetage morphologique de textes français avec un arbre de décisions. *Traitement Automatique des Langues*, **36**(1-2), 23–35.

Stickel, G. (1999). Zur Sprachbefindlichkeit der Deutschen: Erste Ergebnisse einer Repräsentativumfrage. In G. Stickel, editor, *Sprache - Sprachwissenschaft Öffentlichkeit*, pages 16–44. De Gruyter, Berlin.

Summers, D. and Rundell, M., editors (1995). *Longman Dictionary of Contemporary English*. Longman, Harlow, 3rd edition.

Tattersall, A. (2003). The Internet and the French language. Occasional paper. Centre for Language in Education, University of Southhampton, Southampton, UK.

Tautenhahn, K. (1998). *Anglizismen im Sportteil der "Freien Presse" 1985 und 1995. Eine Untersuchung*. Technische Universität Chemnitz, Chemnitz. Dissertation.

All About Market Research (2007). Internet usage growth. In *All About Market Research*. Published on 19/03/2007. Available online at: `http://www.allaboutmarketresearch.com/internet.htm`.

Thogmartin, C. (1984). Some "English" words in French. *The French Review*, **57**(4), 447–455.

Thompson, H. S., Tobin, R., and McKelvie, D. (1997). *LT XML. Software API and toolkit for XML processing*. Available online at: `http://www.ltg.ed.ac.uk/software/`.

Tomaschett, G. (2005). *Anglizismen - Ist die deutsche Sprache gefährdet? Zunahme der Anglizismen in den Inseraten der Schweizer Zeitungen Bote der Urschweiz und Weltwoche bzw. NZZ am Sonntag von 1989 - 2005*. Universität Zürich, Zürich. Ph.D. Thesis.

Tosi, A. (2001). *Language and Society in a Changing Italy*. Multilingual Matters, Toronto.

Traber, C., Huber, K., Jantzen, V., Nedir, K., Pfister, B., Keller, E., and Zellner, B. (1999). From multilingual to polyglot speech synthesis. In *Proceedings of 6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, pages 835–838, Budapest, Hungary.

Trancoso, I., Viana, C., Mascarenhas, I., and Teixeira, C. (1999). On deriving rules for nativised pronunciation in navigation queries. In *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, pages 195–198, Budapest, Hungary.

Turunen, M. and Hakulinen, J. (2000). Mailman - a multilingual speech-only e-mail client based on an adaptive speech application framework. In *Proceedings of the Workshop on Multi-Lingual Speech Communication*, pages 7–12, Kyoto, Japan.

Ustinova, I. P. (2006). English and emerging advertisement in Russia. *World Englishes*, **25**(2), 267–277.

Utzig, T. (2002). *Anglizismen in den Stellenanzeigen der Süddeutschen Zeitung und der Frankfurter Allgemeinen Zeitung*. Tectum, Marburg.

van Rooij, J. C. G. M. and Plomp, R. (1991). The effect of linguistic entropy on speech perception in noise in young and elderly listeners. *The Journal of the Acoustical Society of America*, **90**(6), 2985–2991.

Viereck, W. (1980). Empirische Untersuchungen insbesondere zum Verständnis und Gebrauch von Anglizismen im Deutschen. In W. Viereck, editor, *Studien zum Einfluß der englischen Sprache auf das Deutsche*, pages 237–322. Tübingen, Gunter Narr.

Viereck, W. (1984). Britisches Englisch und amerikanisches Englisch/Deutsch. In *Sprachgeschichte: Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*. Walter de Gruyter, Berlin.

Volk, M. (2001). Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceedings of the International Conference on Corpus Linguistics (CL2001)*, Lancaster, UK.

von Wickert, P. (2001). Acceptance of anglicism in the German advertising. Term Paper, Culture and Social Studies, Hamburg University of Applied Science, Hamburg, Germany.

Wangensteen, B. (2002). Nettbasert nyordsinnsamling. *Språknytt*, **2**, 17–19.

Waterman, J. T. (1991). *A History of the German Language*. Revised edition. Waveland Press, Prospects Heights, Illinois.

Weinrich, H. (1984). Die Zukunft der deutschen Sprache. In B. Carstensen, F. Debus, H. Henne, P. von Polenz, D. Stellmacher, and H. Weinrich, editors, *Die deutsche Sprache der Gegenwart*, pages 83–108. Vandenhoeck & Ruprecht, Göttingen.

Weiss, O. (2005). Security-Tool verhindert, dass Hacker über Google Sicherheitslücken finden. *Computerwelt*. Published on 10/01/2005. Available online at: `http://www.computerwelt.at/detailArticle.asp?a=88302&n=4`.

Wijngaarden, S. J. v. and Steeneken, H. J. (2000). The intelligibility of German and English speech to Dutch listeners. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, volume 3, pages 929–932, Beijing, China.

Yang, W. (1990). *Anglizismen im Deutschen: am Beispiel des Nachrichtenmagazins Der Spiegel*. Niemeyer, Tübingen.

Yeandle, D. (2001). Types of borrowing of Anglo-American computing terminology in German. In M. C. Davies, J. L. Flood, and D. N. Yeandle, 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, Stuttgart.

Yoneoka, J. (2005). The striking similarity between Korean and Japanese English vocabulary - historical and linguistic relationships -. *ASIAN ENGLISHES*, **8**(1), 26–47.

Zhu, X. and Rosenfeld, R. (2001). Improving trigram language modeling with the World Wide Web. In *Proceedings of the International Conference on Acoustic Speech and Signal Processing (ICASSP 2001)*, pages 533–536, Salt Lake City, Utah, USA.

Zimmer, D. E. (1997). *Deutsch und anders. Die Sprache im Modernisierungsfieber*. Rowohlt, Reinbek.