Using Foreign Inclusion Detection to Improve Parsing Performance

Beatrice Alex, Amit Dubey and Frank Keller
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW, UK
{balex,adubey,keller}@inf.ed.ac.uk

Abstract

Inclusions from other languages can be a significant source of errors for monolingual parsers. We show this for English inclusions, which are sufficiently frequent to present a problem when parsing German. We describe an annotation-free approach for accurately detecting such inclusions, and develop two methods for interfacing this approach with a state-of-the-art parser for German. An evaluation on the TIGER corpus shows that our inclusion entity model achieves a performance gain of 4.3 points in F-score over a baseline of no inclusion detection, and even outperforms a parser with access to gold standard part-of-speech tags.

1 Introduction

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 text-to-speech (TTS) literature (see Section 2), 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.

In this paper, we address this challenge. We focus on English inclusions in 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 text can be English inclusions. Even in regular newspaper text as used for many NLP applications, English inclusions can be found in up to 7.4% of all sentences (see Section 3 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 specialized 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 problems with inclusions – most of the time, they are unknown words, and as they originate from another language, standard methods for unknown words guessing (suffix stripping, etc.) are unlikely to be successful. Furthermore, the fact that inclusions are often multiword expressions (e.g., named entities) means that simply part-of-speech (POS) tagging them accurately is not sufficient: if the parser posits a phrase boundary within an inclusion this is likely to severely decrease parsing accuracy.

In this paper, we focus on the impact of English inclusions on the parsing of German text. We describe an annotation-free method that accurately recognizes English inclusions, and demonstrate that inclusion detection improves the performance of a state-of-the-art parser for German. We show that the way of interfacing the inclusion detection and the parser is crucial, and propose a method for modifying the underlying probabilistic grammar in order to

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 151–160, Prague, June 2007.

enable the parser to process inclusions accurately.

This paper is organized as follows. We review related work in Section 2, and present the English inclusion classifier in Section 3. Section 4 describes our results on interfacing inclusion detection with parsing, and Section 5 presents an error analysis. Discussion and conclusion follow in Section 6.

2 Related Work

Previous work on inclusion detection exists in the TTS literature. Here, the aim is to design a system that recognizes foreign inclusions on the word and sentence level and functions at the front-end to a polyglot TTS synthesizer. Pfister and Romsdorfer (2003) propose morpho-syntactic analysis combined with lexicon lookup to identify foreign words in mixed-lingual text. While they state that their system is precise at detecting the language of tokens and determining the sentence structure, it is not evaluated on real mixed-lingual text. A further approach to inclusion detection is that of Marcadet et. al (2005). They present experiments with a dictionary-driven transformation-based learning method and a corpus-based n-gram approach and show that a combination of both methods yields the best results. Evaluated on three mixed-lingual test sets in different languages, the combined approach yields word-based language identification error rates (i.e. the percentage of tokens for which the language is identified incorrectly) of 0.78% on the French data, 1.33% on the German data and 0.84% on the Spanish data. Consisting of 50 sentences or less for each language, their test sets are very small and appear to be selected specifically for evaluation purposes. It would therefore be interesting to determine the system's performance on random and unseen data and examine how it scales up to larger data sets.

Andersen (2005), noting the importance of recognizing anglicisms to lexicographers, tests algorithms based on lexicon lookup, character n-grams and regular expressions and a combination thereof to automatically extract anglicisms in Norwegian text. On a 10,000 word subset of the neologism archive (Wangensteen, 2002), the best method of combining character n-grams and regular expression matching yields an accuracy of 96.32% and an F-score of 59.4 (P = 75.8%, R = 48.8%). This result is unsurprisingly low as no differentiation is made between full-word anglicisms and tokens with mixed-lingual morphemes in the gold standard.

In the context of parsing, Forst and Kaplan (2006) have observed that the failure to properly deal with foreign inclusions is detrimental to a parser's accuracy. However, they do not substantiate this claim using numeric results.

3 English Inclusion Detection

Previous work reported by Alex (2006; 2005) has focused on devising a classifier that detects anglicisms and other English inclusions in text written in other languages, namely German and French. This inclusion classifier is based on a lexicon and search engine lookup as well as a post-processing step.

The lexicon lookup is performed for tokens tagged as noun (*NN*), named entity (*NE*), foreign material (*FM*) or adjective (*ADJA/ADJD*) using the German and English CELEX lexicons. Tokens only found in the English lexicon are classified as English. Tokens found in neither lexicon are passed to the search engine module. Tokens found in both databases are classified by the post-processing module. 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 (Alex, 2005). This score is determined by weighting the number of hits, i.e. the "absolute frequency" by the estimated size of the accessible Web corpus for that language (Alex, 2006). Finally, the rule-based post-processing module classifies single-character tokens and resolves language classification ambiguities for interlingual homographs, English function words, names of currencies and units of measurement. A further post-processing step relates language information between abbreviations or acronyms and their definitions in combination with an abbreviation extraction algorithm (Schwartz and Hearst, 2003). Finally, a set of rules disambiguates English inclusions from person names (Alex, 2006).

For German, the classifier has been evaluated on test sets in three different domains: newspaper articles, selected from the Frankfurter Allgemeine Zeitung, on internet and telecoms, space travel and European Union related topics. Table 1 presents an

Domain	EI tokens	EI types	EI TTR	Accuracy	Precision	Recall	F
Internet	6.4%	5.9%	0.25	98.13%	91.58%	78.92%	84.78
Space	2.8%	3.5%	0.33	98.97%	84.02%	85.31%	84.66
EU	1.1%	2.1%	0.50	99.65%	82.16%	87.36%	84.68

Table 1: English inclusion (EI) token and type statistics, EI type-token-ratios (TTR) as well as accuracy, precision, recall and F-scores for the unseen German test sets.

overview of the percentages of English inclusion tokens and types within the gold standard annotation of each test set, and illustrates how well the English inclusion classifier is able to detect them in terms of F-score. The figures show that the frequency of English inclusions varies considerably depending on the domain but that the classifier is able to detect them equally well with an F-score approaching 85 for each domain.

The recognition of English inclusions bears similarity to classification tasks such as named entity recognition, for which various machine learning (ML) techniques have proved successful. In order to compare the performance of the English inclusion classifier against a trained ML classifier, we pooled the annotated English inclusion evaluation data for all three domains. As the English inclusion classifier does not rely on annotated data, it can be tested and evaluated once for the entire corpus. The ML classifier used for this experiment is a conditional Markov model tagger which is designed for, and proved successful in, named entity recognition in newspaper and biomedical text (Klein et al., 2003; Finkel et al., 2005). It can be trained to perform similar information extraction tasks such as English inclusion detection. To determine the tagger's performance over the entire set and to investigate the effect of the amount of annotated training data available, a 10-fold cross-validation test was conducted whereby increasing sub-parts of the training data are provided when testing on each fold. The resulting learning curves in Figure 1 show that the English inclusion classifier has an advantage over the supervised ML approach, despite the fact the latter requires expensive hand-annotated data. A large training set of 80,000 tokens is required to yield a performance that approximates that of our annotation-free inclusion classifier. This system has been shown to perform similarly well on unseen texts in different domains, plus it is easily extendable to a new language (Alex, 2006).

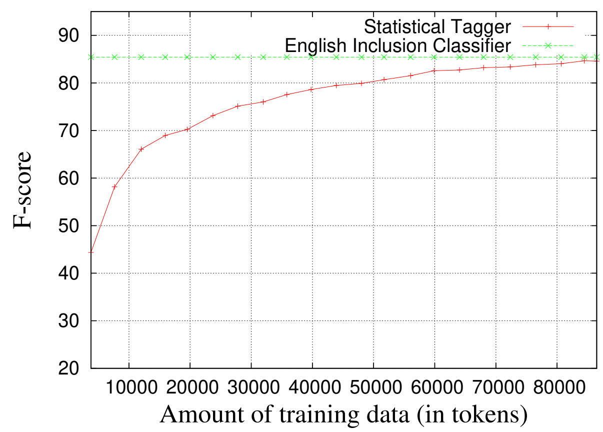


Figure 1: Learning curve of a ML classifier versus the English inclusion classifier's performance.

4 Experiments

The primary focus of this paper is to apply the English inclusion classifier to the German TIGER treebank (Brants et al., 2002) and to evaluate the classifier on a standard NLP task, namely parsing. The aim is to investigate the occurrence of English inclusions in more general newspaper text, and to examine if the detection of English inclusions can improve parsing performance.

The TIGER treebank is a bracketed corpus consisting of 40,020 sentences of newspaper text. The English inclusion classifier was run once over the entire TIGER corpus. In total, the system detected English inclusions in 2,948 of 40,020 sentences (7.4%), 596 of which contained at least one multi-word inclusion. This subset of 596 sentences is the focus of the work reported in the remainder of this paper, 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 illustrated in Figure 2. The tree is relatively flat, which is a trait trait of TIGER treebank annotation (Brants et al., 2002). The non-terminal nodes of the tree represent the phrase categories, and the edge labels the grammatical functions. 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 ot type proper noun component (*PNC*).

4.1 Data

Two different data sets are used in the experiments: (1) the inclusion set, i.e., the sentences containing multi-word English inclusions recognized by the inclusion classifier, and (2) a stratified sample of sentences randomly extracted from the TIGER corpus, with strata for different sentence lengths. 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 the inclusion set has a higher average sentence length than a random sample of sentences from TIGER, and because parsing accuracy is correlated with sentence length. Both the inclusion set and the random set consist of 596 sentences and do not overlap.

4.2 Parser

The parsing experiments were performed with a state-of-the-art 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 unlexicalized probabilistic context-free grammar (PCFG) and relies on 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 parser also makes use of suffix analysis. However, beam search or smoothing are not employed. Based upon an evaluation on the NEGRA treebank (Skut et al., 1998), using a 90%-5%-5% training-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). These figures were derived on a test set limited to sentences containing 40 tokens or less. In the data set used in this paper, however, sentence length is not limited. Moreover, the average sentence length of our test sets is considerably higher than that of the NEGRA test set. Consequently, a slightly lower performance and/or coverage is anticipated, albeit 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 in Brants et. al (2002).

4.3 Parser Modifications

We test several variations of the parser. The **baseline** parser does not treat foreign inclusions in any special way: the parser attempts to guess the POS tag and grammatical function labels of the word 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 properly. Therefore, a suitable upper bound is to give the parser **perfect tagging** information. Two further versions interface with our inclusion classifier and treat words marked as inclusions differently from native words. The first version does so on a **word-by-word** basis. In contrast, the **inclusion entity** approach attempts to group inclusions, even if a grouping is not posited by phrase structure rules. We now describe each version in more detail.

In the TIGER annotation, preterminals include both POS tags and grammatical function labels. For example, rather than a preterminal node having the category *PRELS* (personal pronoun), it is given the category *PRELS-OA* (accusative personal pronoun). Due to these grammatical function tags, the perfect tagging parser may disambiguate more syntactic information than provided with POS tags alone. Therefore, to make this model more realistic, the parser is required to guess grammatical functions (allowing it to, for example, mistakenly tag an accusative pronoun as nominative, dative or genitive). This gives the parser information about the POS tags of English inclusions (along with other words), but does not give any additional hints about the syntax of the sentence.

The two remaining models both take advantage

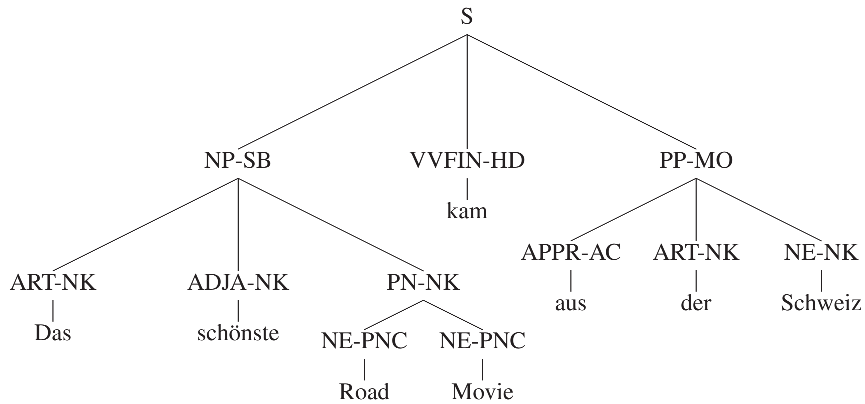
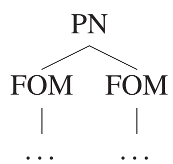


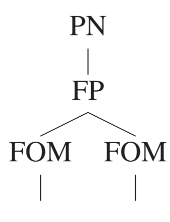
Figure 2: Example parse tree of a German TIGER sentence containing an English inclusion. Translation: The nicest road movie came from Switzerland.

NE	*FM*	*NN*	*KON*	*CARD*	*ADJD*	*APPR*
1185	512	44	8	8	1	1

Table 2: POS tags of foreign inclusions.



(a) Whenever a *FOM* is encountered...



(b) ...a new *FP* category is created

Figure 3: Tree transformation employed in the *inclusion entity* parser.

of information from the inclusion detector. To interface the detector with the parser, we simply mark any inclusion 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 in inclusions during training (the tags listed in Table 2). When a *FOM* is detected, these and only these POS tags are guessed; all other aspects of the parser remain the same.

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 was developed 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 3(a), the tree is modified so that it is the *FP* rule which generates the *FOM* tokens. Figure 3(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 *FOM*s. Once a candidate parse is created during testing, the inverse operation is applied, removing the *FP* node.

4.4 Method

For all experiments reported in this paper, the parser is trained on the TIGER treebank. As the inclusion and random sets are drawn from the whole TIGER 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 multifold cross-validation tests. Using 5 folds, each model is trained on 80% of the data while the remaining 20% are held out. The held out set is then

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	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	62.8	59.0	73.1	99.2	2.1	34.2	70.2
Random set	63.3	67.3	65.2	81.1	99.2	1.6	40.4	75.1
Inclusion entity model								
Inclusion set	61.3	65.9	63.5	78.3	99.0	1.7	42.4	77.1
Random set	63.4	67.5	65.4	80.8	99.2	1.6	40.1	75.7

Table 3: Baseline and perfect tagging for inclusion and random sets and results for the word-by-word and the inclusion entity models.

intersected with the inclusion set (or, respectively, the random set). The evaluation metrics are calculated on this subset 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 the aggregated totals of the five held out test sets. For each experiment, we report parsing performance in terms of the standard PARSEVAL scores (Abney et al., 1991), including coverage (Cov), labeled precision (P) and recall (R), 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 ≤2CB). In addition, we also report dependency accuracy (Dep), calculated using the approach described in Lin (1995), using the head-picking method used by Dubey (2005a). The labeled bracketing figures (P, R and F), and the dependency score are calculated on all sentences, with those which are out-of-coverage getting zero nodes. The crossing bracket scores are calculated only on those sentences which are successfully parsed.

4.5 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 3). However, scores differ for the bracketing measures. Using stratified shuffling[1], we performed a *t*-test on precision and recall, and found both to be significantly worse in the inclusion condition. Overall, the harmonic mean (F) of precision and recall was 65.2 on the random set, 6 points better than 59.2 F observed on the inclusion set. Similarly, dependency and cross-bracketing scores are higher on the random test set. This result strongly indicates 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*), solely associated with inclusions. While perfect tagging increases overall accuracy, a difference of 5.1 in F-score remains between the random and inclusion test sets. Although smaller than that of the baseline runs, this difference shows that even with perfect tagging,

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

parsing English inclusions is harder than parsing monolingual data.

So far, we have shown that the English inclusion classifier is able to detect sentences that are difficult to parse. We have also shown that perfect tagging helps to improve parsing performance but is insufficient when it comes to parsing sentences containing English inclusions. In the next section, we will examine how the knowledge provided by the English inclusion classifier can be exploited to improve parsing performance for such sentences.

4.6 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 of 59.0. All other scores, including dependency accuracy and cross bracketing results are similar to those of the baseline (see Table 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, we report that the difficulty in parsing multiword English inclusions is recognizing them as constituents, rather than recognizing their POS tags. We attempt to overcome this problem with the inclusion entity model.

4.7 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 of 63.5 (4.3 higher) and a dependency accuracy of 78.3% (3.4% higher). The average number of crossing brackets is 1.7 (0.4 lower), with 42.4% of the parsed sentences having no crossing brackets (8.2% 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, F and cross-brackting 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.

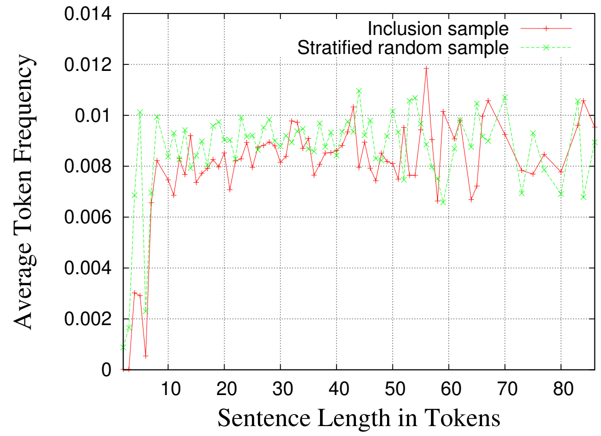


Figure 4: Average relative token frequencies for sentences of equal length.

Not only did the inclusion entity parser perform above the baseline on every metric for the inclusion set, 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 additional information about English inclusions provided by our inclusion classifier. However, it is also important to note that the performance of this model on the inclusion set is still consistently lower than that of all models on the random set. This demonstrates that sentences with inclusions are more difficult to parse than 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 may not contain as many infrequent tokens as the inclusion set. Figure 4 shows the average relative token frequencies for sentences of equal length for both sets. The frequency profiles of the two data sets are broadly similar (the difference in means of both groups is only 0.000676), albeit significantly different according to a paired t-test ($p \leq 0.05$). This is one reason why the inclusion entity model's performance on the inclusion set does not reach the upper limit set by the random sample.

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

Table 4: Gold phrase categories of inclusions.

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

5 Error Analysis

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 recognizing multi-word English inclusions is therefore 95.4% for this TIGER sample.

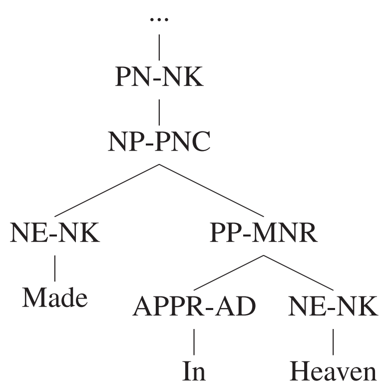
Table 4 illustrates that the majority of multi-word English inclusions are contained in a proper noun (*PN*) phrase, including names of companies, political parties, organizations, films, newspapers, etc. A less frequent phrasal category is chunk (*CH*) which tends to be used for slogans, quotes or expressions like *Made in Germany*. Even in this small sample, annotations of inclusions as either *PN* or *CH*, and not the other, can be misleading. For example, the organization *Friends of the Earth* is annotated as a *PN*, whereas another organization *International Union for the Conservation of Nature* is marked as a *CH* in the gold standard. This suggests that the annotation guidelines on foreign inclusions could be improved when differentiating between phrase categories containing foreign material.

For the majority of sentences (62%), the baseline model predicts more brackets than are present in the gold standard parse tree (see Table 5). This number decreases by 11% to 51% when parsing with the inclusion entity model. This suggests that the baseline parser does not recognize 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.

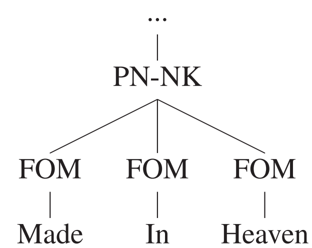
We now turn our attention to how accurately the various parsers are at predicting both phrase bracketing and phrase categories (see Table 6). For 46 (42.2%) of inclusions, the baseline model makes an error with a negative effect on performance. In 39 cases (35.8%), the phrase bracketing and phrase category are incorrect, and constituent boundaries occur within the inclusion, as illustrated in Figure 5(a). Such errors also have 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% inclusions (10.1% more), e.g. see Figure 5(b). 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%). The main confusion lies between assigning the categories *PN*, *CH* and *NP*, the most frequent phrase categories of multi-word English inclusions. This is also partially due to the ambiguity between these phrases in the gold standard. Finally, few parsing errors (4) are caused by the inclusion entity parser due to the markup of false positive inclusions (mainly boundary errors).

6 Discussion and Conclusion

This paper has argued that English inclusions in German text is an increasingly pervasive instance of language mixing. Starting with the hypothesis that such inclusions can be a significant source of errors for monolingual parsers, we found evidence that an unmodified state-of-the-art parser for Ger-



(a) Partial parsing output of the baseline model with a constiuent boundary in the English inclusion.



(b) Partial parsing output of the inclusion entity model with the English inclusion parsed correctly.

Figure 5: Comparing baseline model output to inclusion entity model output.

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

man performs substantially worse on a set of sentences with English inclusions compared to a set of length-matched sentences randomly sampled from the same corpus. The lower performance on the inclusion set persisted even when the parser when given gold standard POS tags in the input.

To overcome the poor accuracy of parsing inclusions, we developed two methods for interfacing the parser with an existing annotation-free inclusion detection system. The first method restricts the POS tags for 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 method yields a performance gain of 4.3 in F-score over a baseline of no inclusion detection, and even outperforms a model involving perfect POS tagging of inclusions.

To summarize, we have shown that foreign inclusions present a problem for a monolingual parser. We also demonstrated that it is insufficient to know where inclusions are or even what their parts of speech are. Parsing performance 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. As our error analysis showed, this prevents cascading errors further up in the parse tree.

Finally, our results indicate that future work could improve parsing performance for inclusions further: we found that parsing the inclusion set is still harder than parsing a randomly sampled test set, even for our best-performing model. This provides an upper bound on the performance we can expect from a parser that uses inclusion detection. Future work will also involve determining the English inclusion classifier's merit when applied to rule-based parsing.

Acknowledgements

This research is supported by grants from the Scottish Enterprise Edinburgh-Stanford Link (R36759), ESRC, and the University of Edinburgh. We would also like to thank Claire Grover for her comments and feedback.

References

Steven Abney, Dan Flickenger, Claudia Gdaniec, Ralph Grishman, Philip Harrison, Donald Hindle, Robert Ingria, Frederick Jelinek, Judith Klavans, Mark Liberman, Mitchell P. Marcus, Salim Roukos, Beatrice Santorini, and Tomek Strzalkowski. 1991. Procedure for quantitatively comparing the syntactic coverage of English grammars. In Ezra Black, editor, *HLT'91: Proceedings of the workshop on Speech and Natural Language*, pages 306–311, Morristown, NJ, USA. Association for Computational Linguistics.

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*, pages 133–138, Ann Arbor, Michigan, USA.

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.

Gisle Andersen. 2005. Assessing algorithms for automatic extraction of Anglicisms in Norwegian texts. In *Corpus Linguistics 2005*, Birmingham, UK.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT02)*, pages 24–41, Sozopol, Bulgaria.

Amit Dubey. 2005a. *Statistical Parsing for German: Modeling syntactic properties and annotation differences*. Ph.D. thesis, Saarland University, Germany.

Amit Dubey. 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.

Jenny Finkel, Shipra Dingare, Christopher D. Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. 2005. Exploring the boundaries: Gene and protein identification in biomedical text. *BMC Bioinformatics*, 6(Suppl 1):S5.

Martin Forst and Ronald M. Kaplan. 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.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 423–430, Saporo, Japan.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-03)*, pages 180–183, Edmonton, Canada.

Dekang Lin. 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.

Jean-Christophe Marcadet, Volker Fischer, and Claire Waast-Richard. 2005. A transformation-based learning approach to language identification for mixed-lingual text-to-speech synthesis. In *Proceedings of Interspeech 2005 - ICSLP*, pages 2249–2252, Lisbon, Portugal.

Beat Pfister and Harald Romsdorfer. 2003. Mixed-lingual analysis for polyglot TTS synthesis. In *Proceedings of Eurospeech 2003*, pages 2037–2040, Geneva, Switzerland.

Ariel Schwartz and Marti Hearst. 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.

Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 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.

Boye Wangensteen. 2002. Nettbasert nyordsinnsamling. *Språknytt*, 2:17–19.