Part-of-Speech Tagging with a Symbolic Full Parser: Using the TIGER Treebank to Evaluate Fips

Yves Scherrer
Language Technology Laboratory (LATL)
University of Geneva
1211 Geneva 4, Switzerland
yves.scherrer@lettres.unige.ch

Abstract
In this paper, we introduce the German version of the multilingual Fips parsing system. We focus on the evaluation of its part-of-speech tagging component with the help of the TIGER treebank. We explain how Fips can be adapted to the tagset used by TIGER and report first results of this study: currently, 87% of words are tagged correctly. We also discuss some common errors and explore a possible extension of this study to parsing.

1 Introduction
Fips is a parsing framework based on the main assumptions of Chomsky’s generative linguistics. It has been designed as a multilingual framework, making it easy to add new languages. Currently, it is available for six languages (English, French, German, Italian, Spanish and Greek). While the French version (providing the best coverage) has taken part in evaluation campaigns (Adda et al., 1998; Goldman et al., 2005), the other language modules have only been subject to internal qualitative evaluation. However, the availability of gold standard treebanks allows for quantitative evaluation of rule-based parsing systems. In particular, we propose to use the TIGER treebank for the evaluation of the German version of Fips.

This paper reports on research in progress. As a preliminary step towards a quantitative assessment of parser performance, we focus on the task of Part-of-Speech (POS) tag comparison here. This task is intended to yield a first appreciation of the quality of the German Fips component without having to deal with the full parser output and its possible incompatibilities due to underlying theoretical differences. Tag comparison operates on a word-by-word basis and provides binary measures of accuracy (tag identity or difference).

We extend our work to the tasks of lemma identification and morphological analysis: Fips as well as the TIGER treebank provide this information.

Fips has been developed independently of the TIGER treebank. Therefore, a large part of this paper deals with problems arising from mismatches between the design decisions made for Fips and the annotation guidelines of TIGER. In our view, a detailed discussion of these mismatches is essential for a fair assessment of the performances of Fips, but may also be interesting for future research involving evaluation.

This paper is organized as follows. In Section 2, we present the Fips framework. In Section 3, we recall the main characteristics of the TIGER treebank, explain the adaptations we applied to the Fips tagger and give some information about the evaluation setup. We go on to report the results for the three main tasks: Part-of-Speech tagging (Section 4), lemma identification (Section 5), and morphological analysis (Section 6). Section 7 compares our work to statistical POS tagging and to parser evaluation. We conclude by giving an overview of the benefits of quantitative evaluation.

2 The Fips framework
Fips (Wehrli, 2007) is a deep symbolic parser developed at the University of Geneva. It currently supports six languages, and others are under development. The parser is based on an adaption of generative linguistics, borrowing concepts from the Minimalist model (Chomsky, 1995), from the Simpler
Syntax model (Culicover and Jackendoff, 2005), as well as from Lexical Functional Grammar (Bresnan, 2001). Each syntactic constituent is represented as a simplified X-bar structure without intermediate levels, in the form \([XP LXR]\). \(X\) denotes a lexical category, \(L\) and \(R\) stand for (possibly empty) lists of left and right subconstituents, respectively.

The originality of Fips lies in its two-layer architecture. Fundamental properties and structures that are common to all languages are defined in an abstract, language-independent layer. On a theoretical level, this layer can be associated to the concept of “universal grammar”. On top of this layer, a particular, language-dependent layer extends the abstract structures and adds language-specific grammar rules. The Fips lexicon contains detailed morphosyntactic and semantic information such as selectional properties, subcategorization information and syntactico-semantic features. The parser is thus based on a strong lexicalist framework. In order to guide ambiguity resolution, numeric penalty values can be assigned to rules and lexemes.

The German component of Fips contains around 100 language-specific grammar rules. The lexicon contains 39 000 lexemes and 410 000 word forms. The word forms are generated by a rule-based morphological generator. The lexicon also contains 500 multi-word expressions and 1500 high-frequency compound nouns. Unknown compound nouns are chunked at runtime.

Fips operates in two modes: parser (see Figure 1) and tagger (see Figure 2) output. The tagger output allows us to benefit from the rich information of the Fips lexicon, being at the same time more robust than the parser.

3 Experimental setup

3.1 The TIGER treebank

The TIGER treebank contains about 50 000 sentences of newspaper text, covering all domains (Brants et al., 2002). The annotation has been performed with the help of interactive tools. This methodology allows the human annotator to easily accept or reject proposals made by the computer. Part-of-speech tags are proposed by a statistical tagger trained on a manually annotated corpus. It uses the Stuttgart-Tübingen-Tagset (STTS) (Thielen et al., 1999). The parse trees were constructed interactively with the help of a statistical parser. Figure 3 shows an example of the TIGER export file.

3.2 Adaptations

In order to compare the Fips output with the TIGER tags, some adaptations had to be made. First of all, the tagset had to be changed to match the STTS tagset. While this procedure was straightforward for most of the categories, it showed that the German tagging module of Fips had never been subject to it in the present study.

Given the scope of this workshop, we forgo translating German examples into English.
to a rigorous evaluation. For example, there were no particular tags for pronominal prepositions (e.g., "darüber, deswegen"), for prepositions with articles (e.g., "beim, ins"), and for the infinitival particle "zu".

Small adaptions concerned the replacement of ß by ss (Fips uses the Swiss Standard German orthography, lacking the letter ß) and the different lemmatization of the particle verbs: in TIGER and in contrast to Fips, the particles are not attached to the lemma (see the verb "andeuten" in Figures 2 and 3).

Finally, the Fips tagger contains a compound noun chunker which is automatically used for unknown words and which outputs one line for each chunk. These lines had to be reassembled to fit with the unchunked TIGER output (cf. the compound noun Obergrenze in Figures 2 and 3).

### 3.3 Evaluation

From the TIGER export file, we extracted the original sentences and submitted them to the Fips tagger. Then, we compared its results with the information given in TIGER. Overall, 792 885 words were compared. This number does not correspond to the 888 578 tokens of the TIGER corpus, because the concept of word is much more flexible in Fips than in TIGER. For example, the token "62jähriger" is split into two words 62 and "jährige". By contrast, vor allem is regarded as a single lexical item (adverb) by Fips, but as two words by TIGER. Moreover, for a currently unknown reason, some words do not show up in the output of the Fips tagger.

### 4 Part-of-speech tagging results

The most important part of this evaluation concerns the part-of-speech tags. As explained above, we have adapted Fips to generate STTS tags. Table 1 shows the number of correctly predicted tags, and

| TIGER Tag | Fips Tag | Number | Percentage |
|-----------|----------|--------|------------|
| NN        | NE       | 12592  | 1.59       |
| KON       | ADV      | 8000   | 1.01       |
| ADJD      | ADV      | 6737   | 0.85       |
| ADV       | PTKA     | 4976   | 0.63       |
| NE        | NN       | 4782   | 0.60       |
| VAFIN     | VVFIN    | 3529   | 0.45       |
| ART       | PRELS    | 2935   | 0.37       |
| VVFIN     | VVIMP    | 1937   | 0.24       |
| VVINF     | VVFIN    | 1859   | 0.23       |
| VVPP      | VVFIN    | 1624   | 0.20       |
| Correct tags |         | 692 386 | 87.32     |
| Tested words |         | 792 885 | 100.00    |

Table 1: Results of the part-of-speech tag comparison. The table shows the number of tags correctly predicted by Fips (second last line), as well as the ten most frequent erroneous predictions. The first column shows the correct tag as given by TIGER, the second column shows the erroneous tag assigned by Fips.
the ten most frequent tagging errors. In the follow-
ing sections, we discuss some of these errors.

4.1 Proper and common nouns

The most common error is related to the distinction between proper (NE) and common nouns (NN). This error affects 2.19% of words (see first and fifth line in Table 1) and accounts for 17.29% of all tagging errors. Currently, the distinction between proper and common nouns is implemented in Fips as follows.

A noun is regarded as common noun if:

- it is present in the lexicon and not explicit-
itly marked as proper noun: Chemie, Hirsch, Konkurrenz, or

- it is a compound noun that can be analyzed into chunks which are present in the lexicon: Bundes+bank, Finanz+markt, Sitz+platz.

A noun is regarded as proper noun if:

- it is explicitly marked as such in the lexicon: Gregor, Berlin, Europa.

- it is not present in the lexicon and cannot be fully analyzed as compound noun: Talk, Gaullismus, Kibbuzarbeiter.

Tagging errors occur in two ways. Words that are annotated as common nouns by TIGER are annotated as proper nouns by Fips (see first line in Ta-
ble 1). This happens for all common nouns that are not present in the lexicon (e.g., Primadonna, Portfolio, Niedersachse, Gaullismus). There are also compound nouns with a proper noun complement: Vichy-Zeiten, Spreearm. While TIGER considers these words as common nouns because the head is a common noun, Fips still analyzes them as proper nouns. For other words like Marseillaise, the TIGER annotation as common noun may be questioned.

In the other way, some TIGER proper nouns have been tagged by Fips as common nouns (cf. fifth line in Table 1). One common category of erroneous tagging is the case of homonymous proper and common nouns. For example, Kohl and Teufel are common nouns, but also the names of German politicians and therefore proper nouns. These misinterpretations are due to the fact that Fips does not contain any specific Named Entity Recognition module. While Fips successfully relies on letter case to identify proper nouns in other languages, this approach obviously does not work in German.

Some proper nouns exhibit a more subtle phe-

nomenon: words like Mannheim, Wendland or Kantstrasse are analyzed by Fips as common compound nouns (Mann+Heim, wenden+Land, Kante+Strasse). Again, a Named Entity Recognition system would prevent such unfortunate analyses. Furthermore, we do not find it compelling to analyze Buddha, Bundesbank and Bundeskriminalamt as proper nouns.

To sum up, the source of noun mistagging is three-
fold. First, the Fips lexicon contains some gaps. Second, the lack of a Named Entity Recognition module in Fips causes an overgeneration of homograph common nouns where a proper noun would be appropriate. Third, the distinction between proper and common nouns is not clear-cut, and some diver-
gences can be considered as normal.

4.2 Conjunctions and adverbs

Conjunctions are frequently mistagged as adverbs. Above all, this error affects the words und, aber, denn, which can have an adverbial (ADV) or a con-
junction (KON) reading. In (1), the first occurrence of und is erroneously tagged as adverb. However, if we parse the first part of the sentence only (2), Fips obtains the correct conjunction reading. This sug-
gests that the conjunction reading is available also for (1), but that the ranking mechanism is flawed and prefers the adverb reading.

(1) Automaten sind dort nur in Geschäften und Restaurants erlaubt und nicht wie in der Bundesrepublik auch im Freien.

(2) Automaten sind dort nur in Geschäften und Restaurants erlaubt.

In general, it seems that Fips gets the conjunctions right in short sentences, while it easily gets confused with longer sentences. However, the preference for the adverbial reading can be easily explained. In or-
der to propose a conjunction, the parser must iden-
tify two conjuncts of the same category, whereas an adverb does not have that requirement. Thus, if the parser fails to find two suitable conjuncts, it will pro-
pose the less constrained adverbial reading.
### 4.3 Adjectives and adverbs

In contrast to English or French, there is no formal difference in German between adjectives used as predicates (e.g., *Er ist schnell*) or as adverbs (e.g., *Er fährt schnell*). This formal identity may have motivated the developers of the STTS tagset to use the same tag (ADJD) in both cases. In contrast, the German Fips tagger is based on earlier work on French and English, where distinct tags for adverbials and predicatives are needed. Therefore, it also uses different tags for German.

We tried to come up with a simple solution to this problem by assigning the ADJD tag to all adverbs whose base forms are homograph with an adjective. However, in this case, we also assigned the ADJD tag to words like *ganz, natürlich, wirklich*, which are tagged as proper adverbs (ADV) in TIGER. In short, we had the choice of either overgenerating ADV tags (keeping the Fips output as-is) or overgenerating ADJD tags (with the homograph modification). Preliminary tests showed similar amounts of overgeneration in both cases. We have thus chosen to stick to the original Fips analyses.

### 4.4 Particles followed by adjectives

STTS introduces a special tag (PTKA) for particles “followed by adjectives or adverbs”, for example *am [schönsten], zu [schnell]*. In Fips, the class of comparative adverbs also contains *auch, so and mehr*. Of course, these words are not always followed by adjectives, and should thus not always be given the PTKA tag. While different readings are indeed available in the Fips lexicon, the results suggest that Fips overgeneralizes the comparative reading and assigns the PTKA tag even in cases where a normal ADV tag would be adequate. (3) shows a sentence where Fips erroneously assigned the PTKA tag to *auch*.

(3) Der Verkehrssenator, wie er künftig auch heißen möge, . . .

### 4.5 Pronouns

The seventh line refers to the homography of the definite determiner and the relative pronoun (PRELS) whenever Fips cannot find an agreement between the determiner and the head of the noun phrase.

(4) Neue Debatte über den Atomschild
In (4), the Fips lexicon only contains the neuter lexeme Schild (which serves as a head of the compound noun Atomschild), but not the rarer masculine homograph lexeme. This lexical gap prevents the masculine determiner den to be attached to Atomschild as a determiner, and Fips resorts to the relative pronoun analysis instead.

4.6 Verb problems

Verb tagging seems to be a serious problem to Fips: four of the ten most frequent tagging errors involve verbs.

The first type of error is related to the distinction between auxiliary and full verbs. The three auxiliary verbs haben, sein, werden can also have full verb readings, depending on the context. We recently observed that Fips preferred the auxiliary reading even in cases where a full verb reading is required, and subsequently modified the constraints on the lexeme selection. It now turns out that these constraints are too strong and lead to a massive overgeneration of the full verb reading.

Then, Fips tends to overgenerate imperatives: third person singular forms are erroneously analyzed as imperative plurals (e.g., kommt, schreit). Again, this is due to agreement constraints: the third person singular requires an overt subject, while an imperative does not. If Fips fails to find a subject that agrees with the verb (for example because of an undetected long distance dependency), it will resort to an imperative reading. In the future development of Fips, further restrictions should be imposed on the use of imperative forms as these are extremely rare in newspaper text.

The last two lines in Table 1 reveal that finite verb forms are preferred to infinite forms: infinitives are mistagged as finite plural forms, and past participles without ge-prefix are mistagged as third person singular forms (for regular verbs) or as past plural form (for irregular verbs with -en participle). These phenomena depend on long distance relations and should typically benefit from a full parsing approach like the one used by Fips. Two factors may explain why this is not the case. First, many sentences in which such errors occur could not be parsed completely by Fips; long distance relations are not fully detected in these cases. Second, the implementation of passive and modal sentences is incomplete and lacks some essential constraints on verb form selection.

5 Lemmatizer results

On the whole TIGER corpus (792 885 words), 94.32% of the words (747 855) were correctly lemmatized. Most errors were due to diverging base form choices. This especially holds for pronouns and nominalized adjectives (cf. Figure 4), but also for pronouns. In TIGER, feminine and neuter pronouns always refer to the masculine lemma, whereas Fips separates the genders more strictly: der (Dat.Sg.Fem) refers to the lemma der (Nom.Sg.Masc) in TIGER, but to die (Nom.Sg.Fem) in Fips. Moreover, participles used as adjectives keep the infinitive as base form in Fips, but not in TIGER.

Some lemma errors are due to wrong POS tagging. For instance, we found that Fips overgenerates imperatives. For example, einig is not analyzed as adjective, but as the imperative singular (with elision of final e) of sich einigen; the adjective nötige is analyzed as the imperative singular of nötigen. However, such awkward analyses should be easy to iron out.

Globally, we find that very few errors are directly due to the lemmatizer; most of them are either due to different base forms or to POS tagging errors.

6 Morphology results

After the discussion of the part-of-speech tagger and lemmatizer functionalities of Fips, we now turn to the last functionality, the morphological analyzer. We restricted our evaluation to the words that obtained correct POS tags: if the POS tag is already wrong, it is very likely that the morphology will be wrong as well. Table 2 reports the results of the mor-
Table 2: Results of the morphological analysis. The table presents the numbers of words that have been correctly analyzed by Fips, and the types of errors that occurred. A word can present several mismatch types.

| Type               | Number | Percentage |
|--------------------|--------|------------|
| Number mismatch    | 15617  | 2.26       |
| Case mismatch      | 12420  | 1.79       |
| Gender mismatch    | 8461   | 1.22       |
| Degree mismatch    | 514    | 0.07       |
| Person mismatch    | 108    | 0.02       |
| Correct analysis or no morphology | 665 110 | 96.06 |
| Tested words       | 692 386 | 100.00 |

It may be interesting to compare Fips to a statistical part-of-speech tagger for German. The TnT tagger (Brants, 2000) is based on Hidden Markov Models, and has been trained and tested on the NEGRA corpus (Skut et al., 1997); NEGRA is the predecessor of TIGER and uses the same tagset. Brants (2000) reports an overall accuracy of 96.7%. However, TnT is not directly comparable to Fips for several reasons.

First, we showed that Fips originally used a different tagset, based on different linguistic assumptions than STTS. Those conceptional differences make up a large part of the errors, as has been shown for the distinction between the ADJD and ADV tags. By contrast, TnT has been trained directly over the STTS tagset and should thus not present such errors.

Second, the recurrence of certain error patterns with Fips illustrates the classical problem of manual rule ranking and weighting in rule-based systems.

Third, Fips has been conceived as a parser in the first place, and its tagger functionality should rather be viewed as a by-product. Hence, its algorithms are not optimized for POS tagging. While there may be simpler approaches to obtain high tagging accuracy, the method chosen for Fips seems theoretically more plausible to us.

As we pointed out at the beginning, this tagger evaluation has been started as a first step towards the evaluation of the Fips parser. While POS tagging has the advantage of operating word-by-word and of being rather theory-independent, these two properties do not hold for parsing.

The phrase trees in TIGER are rather flat, while the ones generated by Fips are deeper and closer to recent generative grammar frameworks. We will thus need to define the type of constituents that can be compared. An even bigger issue is the allowance of discontinuous phrases and crossing branches in TIGER, whereas Fips resolves these phenomena by resorting to projections and traces. Further research has to show if these structural differences can be overcome in order to lead to a meaningful comparison. The exact evaluation metric will also have to be chosen. While PARSEVAL (Black et al., 1991) is still one of the most important metrics, other measures may be more adapted to our problem (Carroll et al., 2002; Rehbein and van Genabith, 2007).

(5) . . . um noch tiefer in den Kosmos blicken zu können.
8 Conclusion

As we remarked above, this article reports on work in progress. Until now, we have been able to show that the general approach of evaluating Fips with the help of the TIGER treebank is valid. With very little adaptation work (see Section 3.2), we managed to obtain 87.32% of POS-tagging accuracy. This is a very promising beginning, and the discussion of the errors has shown that there are many “low hanging fruits” to improve the performance.

In any way, we find that the quantitative evaluation of NLP systems can be quite rewarding: developing rule-based systems is a complex task, often guided by vague intuitions about parsing quality. Quantitative evaluation allows us to measure the progress of the development and guarantees us that improvements on one parameter do not yield unwanted side effects on another.

Finally, the quantitative evaluation of the POS tagging performances yields important feedback on the forces and weaknesses of Fips. The result of the evaluation can be viewed as a sort of priority list for the developer. By working on the most common errors in a target-oriented way, (s)he is guaranteed to invest his/her time in a maximally effective manner. Such guiding principles are very valuable for the further development of any rule-based parsing system, independently of the precise accuracy figures of the evaluation. Even if the adaptation of two different tagsets and tagging philosophies is not straightforward, we plan to extend our evaluation to other languages of the Fips project for which suitable gold standard corpora exist.

Acknowledgements

We thank Eric Wehrli and for his precious support for this work and for his valuable comments on previous versions of this paper.

References

G. Adda, J. Mariani, J. Lecomte, P. Paroubek, and M. Rajmank. 1998. The GRACE French part-of-speech tagging evaluation task. In Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada.

E. Black, S. Abney, S. Flickenger, C. Gdaniec, C. Grishman, P. Harrison, D. Hindle, R. Ingría, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. Procedure for quantitatively comparing the syntactic coverage of English grammars. In HLT ’91: Proceedings of the Workshop on Speech and Natural Language, pages 306–311, Pacific Grove, California.

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

T. Brants. 2000. TnT – a statistical part-of-speech tagger. In Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000), Seattle.

J. Bresnan. 2001. Lexical Functional Syntax. Blackwell, Oxford.

J. Carroll, A. Frank, D. Lin, D. Prescher, and H. Uszkoreit. 2002. Beyond PARSEVAL – towards improved evaluation measures for parsing systems. In Proceedings of the LREC 2002 Workshop, Las Palmas, Gran Canaria.

N. Chomsky. 1995. The Minimalist Program. MIT Press, Cambridge, Mass.

P. W. Culicover and R. Jackendoff. 2005. Simpler Syntax. Oxford University Press, Oxford.

J.-P. Goldman, C. Laenzlinger, G. Soare, and E. Wehrli. 2005. L’analyseur syntaxique multilingue Fips dans la campagne EASy. In Proceedings of TALN XII, volume 2, pages 35–49, Dourdan.

I. Rehbein and J. van Genabith. 2007. Treebank annotation schemes and parser evaluation for German. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL 2007), pages 630–639, Prague.

W. Skut, B. Krenn, T. Brants, and H. Uszkoreit. 1997. An annotation scheme for free word order languages. In Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97, Washington, DC.

C. Thielen, A. Schiller, S. Teufel, and C. Stöckert. 1999. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, University of Stuttgart and University of Tübingen.

E. Wehrli. 2007. Fips, a “deep” linguistic multilingual parser. In Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing, pages 120–127, Prague.