BEYOND IDENTITY: THE CASE OF A COMPLEX HUNGARIAN REFLEXIVE

György Rákosi

University of Debrecen

Proceedings of the LFG09 Conference

Miriam Butt and Tracy Holloway King
(Editors)

2009

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

It is a well-known typological universal that long distance reflexives are generally monomorphemic and complex reflexives tend to be licensed only locally. I argue in this paper that the Hungarian body part reflexive *maga* 'himself' and its more complex counterpart *önmaga* 'himself, his own self' represent a non-isolated pattern that adds a new dimension to this typology. Nominal modification of a highly grammaticalized body part reflexive may reactivate the dormant underlying possessive structure, thereby granting the more complex reflexive variant an increased level of referentiality and syntactic freedom. In particular, the reactivation of the possessive structure in *önmaga* is shown to be concomitant with the possibility of referring to representations of the self, as well as a preference for what appears to be coreferential readings and the loss or dispreference of bound-variable readings.

1. Introduction

According to an established typology, complex reflexives are expected to be local and relatively well-behaved from a binding theoretical perspective, whereas long distance reflexives tend to be monomorphemic (see Faltz 1985, Pica 1987 and subsequent work, as well as Dalrymple 1993 and Bresnan 2001 in the LFG literature). Polymorphemic reflexives, however, are not uniform as they may show different types of morphological complexity. In particular, body part reflexives, which owe their complexity to their historical origin as possessive structures, are often grammatical outside of the local domain in which their antecedent is located. This is a prima facie problem for the typology, since long distance reflexives are not expected to be morphologically complex.

The existence of long distance uses of body part reflexives can be explained under the assumption that these reflexives have a syntactically active possessive structure. Kornfilt (2001) argues that it is exactly such a structure that licenses the Turkish *kendisi* 'himself' both in local and non-local contexts. But this assumption is not necessary, and others have rejected the possessive analysis of non-strictly local body part reflexives (see, for example, the analysis Beavers & Koontz-Garboden 2006 put forward for the English colloquial reflexive *his ass*).

In this paper, I bring evidence from Hungarian to argue for a constrained application of the possessive analysis to complex body part reflexives. The primary reflexive strategy in Hungarian involves the use of the highly grammaticalized body part reflexive *maga* 'himself', which has a more complex variant *önmaga* 'himself, his own self'. I will argue that only the more complex *önmaga* can project a possessive structure in one of its two uses, in essence reactivating an underlying structure that appears to have been lost during the grammaticalization of the primary reflexive. The possessive reanalysis correlates with changes in the syntax and semantics of the complex anaphor *önmaga*. In particular, the reflexive becomes grammatical as a subject and it shows invariable 3SG agreement. Bound variable readings are lost or are dispreferred, and the reflexive can refer to representations of the self, rather than encoding true identity with the referent of the antecedent.

I will use this analysis to argue that on closer inspection, complex body part reflexives which allow for long distance uses do not refute Faltz's (1985) typology. They simply fall outside of the scope of this typology and in fact add a new dimension to it.

The structure of the paper is as follows. In section 2, I give a brief summary of how morphological complexity is known to interact with the size of the reflexive binding domain, paying special attention to body part reflexives. In section 3, I describe the morphology of the two Hungarian reflexives discussed here, and briefly overview the available literature. In section 4, I show that *önmaga* 'himself, his own self' is less constrained syntactically than the primary reflexive *maga* 'himself', but this cannot be explained by simply analyzing *önmaga* as an emphatic reflexive element. This paves the way for a presentation of the peculiar syntactic and semantic properties of *önmaga* in section 5. I conclude in section 6 by showing how the possessive analysis can account for the observed properties of *önmaga*, and round up in section 7 with a cross-linguistic outlook on the implications of the current analysis.

2. Complex reflexives

In his thorough typological survey of reflexives, Faltz (1985) distinguishes between *pronominal* and *compound* (here: complex) reflexives. The third person Norwegian *seg,* the German *sich* or the Russian *sebja* are representatives of the first strategy. The second strategy consists of two broader morphological types. What Faltz calls adjunct reflexives are complexes of a pronoun plus an emphatic marker, like the English *himself* or the Norwegian *seg selv* 'himself'. The other major group consists of body part reflexives (or head reflexives in Faltz's terminology), which start their historical development as a possessive structure and can then become grammaticalized to differing degrees. The Basque *bere burua*, for example, is still ambiguous between the readings 'himself' and 'his head' (Faltz 1985: 32).

The fundamental typology on the correlation between the morphological form of reflexives and their domain of licensing consists of two partially independent statements (see Faltz 1985, Pica 1987, and Cole et al. eds. 2001, among others, and Dalrymple 1993 and Bresnan 2001 in LFG):[1]

(1) **Complex reflexive typology**
a. Long distance reflexives are monomorphemic.
b. Complex reflexives need local antecedents.[2]

Despite occasional skepticism (cf. Büring: 2005, fn. 37), the typology does seem to be making good predictions for adjunct reflexives. The following Norwegian data serve to illustrate the point (Bresnan 2001: 284):

(2) a. *Ola overgår seg selv/*seg.*
Ola surpasses
'Ola surpasses himself.'

b. *Ola bad oss snakke om *seg selv/seg.*
Ola asked us talk.INF about
'Ola$_i$ asked us to talk about him$_i$.'

Seg selv is a complex reflexives and is only licensed locally, and only *seg* can be used as a long distance reflexive form.[3]

Interestingly, reported instances of complex reflexives that do not obey (1b) since they are grammatical both with local and non-local antecedents are all body part reflexives. Let me mention here three such reflexive forms.

The first is the colloquial English *his ass*, discussed by Beavers & Koontz-Garboden (2006).[4] They argue that this reflexive form is a universal pronoun in the sense of Kiparsky (2002), that is, it is grammatical with local ((3a)) as well as non-local antecedents ((3b)), and may even pick up its referent deictically from discourse ((3c)):

[1] The typology covers the default case, in which the reflexive receives no special prosodic prominence and the (local) licensing predicate is other-oriented.

[2] I assume that the local domain relevant for binding theory is defined by the notion of Minimal Complete Nucleus (cf. Dalrymple 1993 and Bresnan 2001): the antecedent of the anaphor must be in the smallest f-structure that contains the f-structure of the anaphor and a SUBJ function. This will suffice for the purposes of this paper.

[3] *Seg* also has nuclear uses; see Lødrup (2007) for details.

[4] I thank the participants of the Cambridge LFG conference for calling my attention to this article.

(3) a. *But most people do believe OJ bought his ass out of jailtime.*
b. *The more he whined about it, the more they nailed his ass.*
c. *I mean her ass, over there.*

The problematic cases for the typology are (3b) and (3c), since they involve the complex reflexive taking a non-local antecedent.

Turkish represents an even more intriguing case (Kornfilt 2001, and also Faltz 1985 and Enç 1987). The primary reflexive *kendi* is a body part reflexive. Its paradigm includes inflected first and second person forms, as well as a non-inflected third person form in the singular and the plural, all of which are local ((4a)). The bare third person singular form *kendi* contrasts with the inflected third person form *kendisi* (and similarly in third person plural), which can take local or long distance antecedents, and even discourse antecedents ((4b)). The examples are from Kornfilt (2001: 198).

(4) a. *Fatma* [*Ahmet-nin kendi-i çok beğen-diğ-in*]-*i biliyor*.
Fatma Ahmet-GEN self-ACC very admire-GER.3SG-ACC knows
'Fatma$_i$ knows that Ahmet$_j$ admires self $_{*i/j/*k}$ very much.'

b. *Fatma* [*Ahmet-nin kendi-sin-i çok beğen-diğ-in*]-*i biliyor*.
Fatma Ahmet-GEN self-3SG-ACC very admire-GER.3SG-ACC knows
'Fatma$_i$ knows that Ahmet$_j$ admires self $_{i/j/k}$ very much.'

What escape the typology in (1b) are the inflected third person reflexives, which can, but need not, take local antecedents, in contrast to inflected first and second person reflexives and non-inflected third person reflexives, which are only locall licensed, as expected.

A third problem case is the Chinese (Mandarin) *ziji* - *ta ziji* 'himself' pair. Presumably, *ziji* might be derived from the meaning 'nose', though this etymology is debatable (Huba Bartos p.c., and see also König & Gast 2006: 264). *Ta* is the third person singular pronoun. Whether *ziji* is a body part reflexive or not, it allows for long distance uses, and, interestingly, *ta ziji* does the same. Pan (1998: 775-76) actually reports that if the antecedent does not c-command the reflexive, he finds *ta-ziji* better than *ziji*.

(5) [*Zhangsan de jiao'ao*] *haile ziji* / *ta-ziji*.
Zhagsan gen pride hurt.PERF self
'Zhangsan$_i$'s pride hurt him$_i$.'

That *ta-ziji* is thus a problem for the complex reflexive typology is also mentioned in Bresnan (2001: 301).

Summing up, body part reflexives may represent a general problem for the typology in (1), but what is especially troubling is the existence of the Turkish and Chinese reflexive pairs. The reflexive typology appears to suggest that increasing the morphological complexity of a reflexive will decrease the size of its binding domain. This does not happen in Chinese, since both *ziji* and *ta ziji* have roughly the same distribution, involving long distance uses. And in Turkish, the morphologically more complex inflected reflexive (*kendisi*) has a wider distribution than the local non-inflected reflexive (*kendi*).

As we will see, Hungarian repeats the Turkish pattern, and thus represents another challenge. But this, as I intend to show here, is only apparent once we realize that we are dealing here with a phenomenon that is simply not covered by the typology in (1).

3. Hungarian reflexives: the background

3.1. The morphology of the two Hungarian reflexives

The primary Hungarian reflexive, which has roughly the same distribution as the English *himself*, is *maga*. The stem is reconstructed to have been used as a word for *body*, but this meaning was lost long ago and in fact native speakers do not have the intuition that the reflexive is compositional. *Mag* in current Hungarian means ‘seed’.

However, the reflexive still shows signs of its possessive origin and it bears possessive type agreement morphology. In Table 1 below, I compare the possessive paradigm of *maga* ‘himself’ and *magja* ‘his seed’.[5] The latter represents the productive morphological pattern, and boldface is used to mark the places where the productive pattern differs from the paradigm of the reflexive. There are two important points of divergence. First, the definite article is obligatory in the possessive construction if the possessor is a (pro-dropped) pronoun, but the reflexive *maga* does not co-occur with the definite article. Second, the phonological shape of the inflectional morphology is not identical in the two paradigms.

[5] Given that Hungarian is a *pro*-drop language, pronominal possessors are normally not pronounced. Note also that Hungarian does not have grammatical gender, so third person pronouns do not manifest gender-related variation in form.

	maga 'HIMSELF'	POSSESSIVE PARADIGM
1SG	*magam* 'myself' mag.1SG	***a*** *mag-**om*** 'my seed' the seed.1SG.POSS
2SG	*magad* 'yourself' mag.2SG	***a*** *mag-**od*** 'your seed' the seed-2SG.POSS
3SG	*maga* 'himself' mag.3SG	***a*** *mag-**ja*** 'his seed' the seed-3SG.POSS
1PL	*magunk* 'ourselves' mag.1PL	***a*** *mag-unk* 'our seed' the seed-1PL.POSS
2PL	*magatok* 'yourselves' mag.2PL	***a*** *mag-**otok*** 'your seed' the seed-2PL.POSS
3PL	*maguk* 'themselves' mag.3PL	***a*** *mag-**juk*** 'their seed' the seed-3PL.POSS

Table 1.

It is the possessive paradigm that has the productive morphophonology, which is a clear indication that the reflexive is highly grammaticalized. Nevertheless, it is also evident that both paradigms utilize the same type of agreement morphology.

Önmaga is the complex of the primary reflexive *maga* and the nominal prefix *ön-* 'self'. This prefix, much like its English counterpart, normally combines with deverbal nouns ((6a)) or participles ((6b)), but it can also be attached to simple, non-eventive nouns ((6c)).

(6) a. *ön-ellát-ás*
self-serve-NOMINAL.SUFFIX
'self-service'

b. *ön-működ-ő*
self-operate-PARTICIPIAL.SUFFIX
'self-operating'

c. *ön-hiba*
self-fault
'(one's) own fault'

Thus we could draw a formal analogy between the possessive form of (6c) and *önmaga*:

(7) a. *ön-hibá-m*
self-fault-1SG.POSS
'my own fault'

b. *ön-magam*
self-mag.1SG
'my own self'

Note that I am translating here *önmagam* as 'my own self' only as an attempt at illustrating how it might differ from *maga* 'himself' in meaning, but the claim is certainly not that *önmaga* and (the Hungarian for) *his own self* are direct grammatical and semantic equivalents of each other.

3.2. The previous literature on *ömaga* 'his own self'

As has been stated above, the primary reflexive strategy in Hungarian involves the use of *maga* 'himself'. *Önmaga* 'his own self' has received relatively little specific attention in the pertinent syntactic literature. In fact the two reflexives are generally treated as essentially equivalent without further comment, and *önmaga* may even be used to illustrate basic binding data in Hungarian (as happens in É. Kiss 1994: 23-26 or É. Kiss 2002: 35-40).

It does appear at first sight that the two reflexives have the same distribution, roughly similar to that of the English *himself*:

(8) *János* *felismerte* (*ön*)*magá-t* *a* *kép-en*.
John.NOM recognized himself-ACC the picture-on
'John recognized himself in the picture.'

The occasional remark that one may find (especially in the descriptivist literature) is that *önmaga* is more emphatic than *maga*, but the nature of this difference is not spelled out in any detail. The only work which goes beyond this remark is Everaert & Szendrői (2002). They note that only *maga*, but not *önmaga* may form part of idiomatic expressions ((9a)), and that *maga* tends to be adjacent with the verb and bear one accent with it ((9b)). The brackets in (9b) are to be interpreted disjunctively.

(9) a. *János nem izgatja* (**ön*)*magá-t.*
John not excites himself-ACC
'John can't be bothered.'

b. *János megmutatta* (*magá-t*) *Marinak* ($^{?}$*magá-t*).
John showed himself-ACC Mary-DAT himself-ACC
'John showed himself to Mary.'

They conclude that whereas *maga* is a simple NP, the more complex *önmaga* projects an extended nominal phrase, or DP.

Though I believe the analysis that Everaert & Szendrői (2002) offer is a step towards a better understanding of the difference between the two Hungarian reflexives, it does not account for further peculiar properties of *önmaga*, which I will show to exist. In the rest of the paper, I undertake a detailed investigation of the diverging grammar of *maga* 'himself' and *önmaga* 'his own self', and offer an alternative analysis that I believe to provide an account of the observed differences and that I hope accommodates the Hungarian data within a larger cross-linguistic domain.

4. *Maga* vs *önmaga*: the basics

4.1. *Önmaga* is less constrained

Maga is a nuclear anaphor, and is licensed as such only in the presence of local antecedents. As I noted in the previous section, *önmaga* is also acceptable in the same local binding domain, so the two are often interchangeable from a purely syntactic perspective. There are, however, constructions in which only *önmaga* is grammatical, and *maga* is ruled out. I briefly survey these contexts here.

First, *maga* is normally not grammatical if embedded within obviously non-argument expressions like the high-level adjunct in (10a) or the passive *by*-phrase with the participle in (10b). *Önmaga*, however, is acceptable in the selfsame contexts.

(10) a. *Önmaga* / **maga szerint János okos ember.*
himself.NOM according.to John clever man
'According to himself, John is a clever man.'

b. *az önmaga* / **maga által okos-nak tart-ott ember*
the himself.NOM by clever-DAT consider-PART man.NOM
lit. 'the man who is considered to be clever by himself'

Second, *önmaga* also shows apparent long distance uses, though it sounds best if it occurs adjacent to the clause in which its antecedent is embedded.

(11) *János fél, hogy *(ön)magát sem választ-ják meg.*
John afraid.is that himself-ACC neither elect-3PL PARTICLE
lit. 'John is afraid that they will not elect himself either.'

Such long distance uses generally occur in point-of-view contexts.

Third, it has been noted in the literature that *önmaga*, unlike *maga*, can function as a nominative subject if it is no more prominent thematically than its non-subject antecedent (see Everaert & Szendrői 2002 and Rákosi 2006, as well as É. Kiss 2002, who does not mention though that *maga* is in fact ungrammatical as a subject). This mainly covers object and dative experiencer verbs, like the following:

(12) *János-t meglepte *(ön)maga.*
John-ACC suprised himself.NOM
lit. 'Himself surprised John.'

In section 5.2, I will argue that once the right context is set up, *önmaga* is licensed as a syntactic subject by any predicate. But the immediate point is that *maga* is never acceptable as a syntactic subject.

I should hasten to add that even if *önmaga* is freer syntactically than *maga*, it is not as obviously free as the colloquial English *his ass* or the Turkish *kendisi* 'himself'. Compare (4b), repeated as (13a), with (13b):

(13) a. *Fatma [Ahmet-nin kendi-sin-i çok beğen-diğ-in]-i biliyor.*
Fatma Ahmet-GEN self-3SG-ACC very admire-GER.3SG-ACC knows
'Fatma$_{i}$ knows that Ahmet$_{j}$ admires self$_{i/j/k}$ very much.'

b. *Fatma tudja, hogy Ahmed nagyon szereti önmagá-t.*
Fatma knows that Ahmet very likes himself-ACC
'Fatma$_{i}$ knows that Ahmet$_{j}$ admires self$_{*i/j/*k}$ very much.'

Though the just discussed differences do exist, it still holds that *önmaga* is not an all purpose reflexive. As (13b) demonstrates, *önmaga* does not always allow for long distance uses, and it does not normally take discourse antecedents. Or, to be more precise, it does not do so the same way as *his ass* or *kendisi* do.

Nevertheless, this situation does represent a problem for the complex reflexive typology. *Maga* is a complex body part reflexive, and it behaves as expected since it is a nuclear anaphor. The even more complex *önmaga*, however, is not necessarily nuclear, and has a wider distribution than the primary reflexive.

4.2. *Önmaga* is not an emphatic form

One potential explanation for the less constrained syntax of *önmaga* could be that it is a special emphatic form, and as such, it is not subject to the bounds of binding theory. Cole, Hermon & Lee (2001: 36) offer some arguments for such an account of the Chinese reflexive *ta ziji*. They claim that *ta ziji* can in fact be analyzed as the complex of the pronoun *ta* 'he' and the reflexive *ziji* 'himself' as an intensifying element. In essence, rather than being a complex reflexive pronominal, *ta ziji* would then be equivalent to the English *he himself* (cf. *John said that he himself wanted to do it*).

Irrespective of whether this analysis works for Chinese or not, it clearly cannot be applied to the case of *önmaga*. The Hungarian intensifier is in fact *maga*, complying with the known fact that primary reflexives often function as intensifiers (cf. König & Gast 2006). *Önmaga* cannot or only marginally can associate with noun phrases as an intensifier:

(14) *Maga/*önmaga az elnök beszélt velünk.*
himself the president.NOM talked with.1PL
'The president himself talked to us.'

Neither can *önmaga* substitute for a pronoun + intensifier unit:

(15) *Ez-t* [*neki magá-nak*] / [**önmagá-nak*] *add oda.*
this-ACC DAT.3SG himself-DAT himself-DAT give.IMP.2SG PART
'Give this to him himself.'

So the relative syntactic freedom of *önmaga* cannot be explained by assuming that this reflexive may function as intensifier.

5. *Maga* vs *önmaga*: beyond identity

5.1. *Önmaga* resembles proper nouns

In 4.1, I focused on some of the usual contexts to show that the syntax of *önmaga* is not identical to that of *maga*. It is, however, more revealing than the previous data set that *önmaga*, unlike *maga*, pattern with proper nouns in certain constructions. I discuss here two such constructions.

First, both proper nouns and *önmaga* can be used predicatively in identity statements. Besides, *önmaga* can also be interpreted as what König & Gast (2006) call an *adverbial-exclusive intensifier* (≈'alone').[6] This is the only reading for *maga*, which cannot be the predicate of an identity statement. Compare:

(16) a. *Újra Péter vagyok.*
again Peter am
'I am (the good old) Peter again.'

b. *Újra önmagam vagyok.*
again myself am
(i) 'I am myself again.'
(ii) 'I am alone again.'

c. *Újra magam vagyok.*
again myself am
(i) *'I am myself again.'
(ii) 'I am alone again.'

[6] This is not in contradiction with what I claimed in 4.2., namely that *maga* is the basic intensifier in Hungarian. *Önmaga* is best as an intensifier on the 'alone' reading, when it still needs to be separated from its associate or its associate needs to be *pro*-dropped. *Maga* is subject to no such restrictions on its intensifier use.

Second, proper nouns can be restrictively premodified when the same real-world individual is conceptualized as corresponding to two partially non-identical *self*s. *Önmaga* can also be premodified this way, but *maga* cannot:

(17) a. *a* *Kádár-kor-i* *Péter*
the Kádár-era-ADJECTIVAL.SUFFIX Peter
'the Peter of the Kádár-era'

b. *a* *Kádár-kor-i* *önmagam*
the Kádár-era-ADJECTIVAL.SUFFIX myself
'my Kádár-era self'

c. **a* *Kádár-kor-i* *magam*
the Kádár-era-ADJECTIVAL.SUFFIX myself
intended: 'my Kádár-era self'

Notice that the grammaticality contrast between *maga* and *önmaga* is very sharp both in (16) and in (17).

What these data suggest is that *önmaga*, unlike a regular reflexive pronominal, shows an increased level of referentiality. It cannot be a simple accident that it patterns with proper nouns in the contexts just discussed.

5.2. Representations of the self

I claimed above that *maga* can normally be substituted for *önmaga*. But now that we have reasons to suspect that the two are not equivalent to each other semantically, it is easier to realize that in certain contexts *önmaga* will be the better or the only option even if the antecedent is locally available.

In general, *önmaga* is felt to be more natural when the context is such that it facilitates a reading in which complete semantic identity does not hold between the antecedent and the reflexive. Consider these two sentences:

(18) a. *A* *történelem* *ismétli* $^?$*magá-t* / *önmagá-t*.
the history.NOM repeats itself-ACC
'History repeats itself.'

b. *János* *ellentmond* $^?$*magá-nak* / *önmagá-nak*.
John.NOM contradicts himself-DAT
'John contradicts himself.'

The reflexive relation that the predicates *repeat* and *contradict* encode is a non-trivial one, for one may only repeat or contradict temporally different states of the self. In other words, (18a) asserts that the current state of history is in some sense equivalent to one of its previous states. The semantic relation between the antecedent and the reflexive is not strict identity, and in such cases, *maga* sounds degraded but *önmaga* is perfectly natural.

The difference is stronger in the so-called 'Mme Tussaud' contexts of Jackendoff 1992 (see also Culicover & Jackendoff 2005). (19) is meant to describe an accident upon Ringo's visit to the wax museum.

(19) *Ringo fell on himself.*
(i) 'The actual Ringo fell on the statue of Ringo.'
(ii) *'The statue of Ringo fell on the actual Ringo.'

Jackendoff points out that (19) can only have the reading in which the actual Ringo falls on the statue Ringo, but not vice versa. What is important for us now is that the English reflexive can apparently be used to refer to representations of the self.

The following variety of the 'Mme Tussaud' context is based on Reuland (2001: 483) and serves to illustrate the Hungarian facts:

(20) a. *Ringo megpillantotta magá-t a tükör-ben.*
Ringo caught.sight.of himself-ACC the mirror-in
(i) 'The actual Ringo saw his own image.'
(ii) $^{??}$'The actual Ringo saw the image of his statue.'

b. *Ringo megpillantotta önmagá-t a tükör-ben.*
Ringo caught.sight.of himself-ACC the mirror-in
(i) 'The actual Ringo saw his own image.'
(ii) 'The actual Ringo saw the image of his statue.'

Though there is some variation in judgments, the statue-reading is only licensed with *önmaga* for most speakers, whereas *maga* may only very marginally allow for this reading.

The clearest cases are those when an ontologically independent and fully functioning copy of the self is created. Such contexts are mostly imaginary, but we do have means of talking about strongly intensional worlds. Imagine, for example, that Peter was cloned or he traveled back in time, and walking on the corridor, he met his own copy. To describe this situation, *önmaga* must be used.

(21) *Önmaga jött Péter-rel szembe a folyosó-n.*
himself came.3SG Peter-with against the corridor-on
lit. 'Himself was coming towards Peter in the corridor.'

There are two noteworthy aspects of (21). First, just like in the English example (19), the reflexive must refer to the copy and the proper name refers to the real (i.e., the original) Peter. Second, in these 'representations of the self' contexts *önmaga* is grammatical as a subject by any predicate. I noted in subsection 4.1 that it is known in the literature that *önmaga* can mostly be the subject of experiencer predicates. In the light of (21), we can now interpret this as derivative of the fact that experiencer predicates facilitate at least weak 'representations of the self' readings. If, after all, *John is surprised by himself* is true (cf. 12), then it must be the case that what surprises John is an aspect of his personality that he was not aware of. It seems that this level of conceptual differentiation is enough to license *önmaga* as a subject. With non-experiencer predicates, stronger contextual support is required to achieve the same affect.

There is a further peculiar property of the subject uses of *önmaga*. Irrespective of which form of the paradigm is used, these reflexive subjects will always trigger third person singular agreement on the verb. In (22), the subject is the first person singular reflexive, but the verb is still in its third person singular form.

(22)	*Önmagam*	*jött*	*velem*	*szembe*	*a*	*folyosó-n*.
	myself	came.3SG	with.1SG	against	the	corridor-on

lit. 'Myself was coming towards me in the corridor.'

Given that otherwise agreement is applied across the board in Hungarian, it is strange that now we seemingly face its absence. Notice also that my real self is referred to by the pronominal *velem* 'with me', rather than by an anaphor. That is also unexpected. If *önmagam* 'myself' was a first person singular form, then the coreferring pronominal would have to be ungrammatical in the same clause.

5.3. Two entries for *önmaga*

I concluded the last subsection with an apparent puzzle. *Önmaga* shows the full agreement paradigm (cf. Table 1); still it always triggers third person singular agreement if it is used as a subject. What I want to suggest now is that in fact we have two separate lexical entries for *önmaga* (all through the paradigm). *Önmaga*$_1$ is a more or less regular reflexive, except for the fact that it is not strictly nuclear (4.2). It agrees with its antecedent in person and number, and it cannot be used as a subject. *Önmaga*$_2$ is a special type of reflexive: this is the one that is used in 'representations of the self' contexts. It can be used as a subject, and it shows constant third person singular agreement with the verb.

One intuitively appealing motivation for this move is that this way we can clearly separate relations of true identity (*önmaga*$_1$) from relations of referential differentiation (*önmaga*$_2$). Note that in a sentence like (11), repeated here as (23), the antecedent and the reflexive clearly refer to one single conceptualization of the same individual:

(23) *János fél, hogy önmagát sem választ-ják meg.*
John afraid.is that himself-ACC neither elect-3PL PARTICLE
'John is afraid that they will not elect himself either.'

We can just simply describe this fact by assuming that the sentence contains *önmaga*$_1$.

An argument with more substantial weight is based on patterns of licensing bound variable and coreference readings (see Evans 1980, Reinhart 1983, 2006, Bresnan 2001 and Büring 2005, among others, for this difference). The less complex reflexive *maga* only seems to allow for bound variable readings, but not for coreference readings, as the following VP-ellipsis context testifies:

(24) *János látja magá-t, de Kati nem.*
John see.3SG himself-ACC but Kate.NOM not
(i) 'John sees himself, but Kate (does) not (see herself).'
(ii) *'John sees John, but Kate (does) not (see John).'

(i) is the sloppy, bound variable reading, under which the elided anaphor is understood to be locally bound by the subject of the clause in which the VP is missing. Under the strict, coreference reading (ii), what Kate does not see is John, not herself. If, however, we replace *maga* with *önmaga*, then the coreference reading becomes fully grammatical for many speakers, and marginally available for others.

(25) *János látja önmagá-t, de Kati nem.*
John see.3SG himself-ACC but Kate.NOM not
(i) 'John sees himself, but Kate (does) not (see herself).'
(ii) $^{\surd/??}$'John sees John, but Kate (does) not (see John).'

Notice that the bound variable reading is still available.

And now let us consider (26), where the reflexive is the subject and the antecedent is the object.

(26) *Engem megijeszt önmagam, de téged nem.*
I.ACC scares myself.NOM but you.ACC not
(i) $^{*/??}$'Myself scares me, but (yourself does) not (scare) you.'
(ii) 'Myself scares me, but (myself does) not (scare) you.'

Interestingly, now the bound variable reading (i) becomes unavailable for many speakers, or at least very marginal for others, but the coreference reading (ii) is grammatical. This is in clear contrast with (25).

One convenient way of explaining this contrast is to assume that the entry that we have in (26) is our *önmaga*$_2$, *which does not license bound* *variable readings*. Under this account, *önmaga*$_1$ can be considered to be a regular reflexive that favours bound variable readings. This is the entry we have in (25).

Two remarks need to be added to this. First, one could object that the bound variable reading is unavailable in (26) because the construction is an instance of *weak crossover*. But note that weak crossover effects are not attested in Hungarian as long as the object binder linearly precedes the bound variable in the subject. (27) is identical to (26) in every respect, except for the fact that it has a possessive noun phrase subject:

(27) *Engem szeret az anyá-m, de téged nem.*
I.ACC loves the mother-1SG.POSS but you.ACC not
(i) 'My mother loves me, but (your mother does) not (love) you.'
(ii) 'My mother loves me, but (my mother does) not (love) you.'

Second, it needs to be admitted that the constraint against bound variable readings of *önmaga*$_2$ is valid for instances of VP ellipsis, but not necessarily for cases of binding by a universal quantifier. (28) is somewhat marked, but it is acceptable nevertheless on what appears to be a bound variable reading. Notice that it is the only possible reading anyway.

(28) *Mindenki-t megijeszt önmaga.*
I-ACC scares himself.NOM
'Everybody is scared by himself.'

Nevertheless, the contrast between (25) and (26) is real. I conclude by maintaining that önmaga has two lexical entries. But we need to weaken the claim made above: *önmaga*$_2$ generally *disallows bound variable readings if the coreference reading is otherwise available*. This is clearly not the way a proper reflexive anaphor is expected to behave.

6. The possessive analysis of *önmaga*$_2$

For the sake of comparison, let me start with the proposed LFG-style entry for the reflexive *maga* 'himself'. I assume a standard LFG binding account in defining the *Minimal Complete Nucleus* as the local binding domain (see footnote 2), and in modelling the semantic relation between the antecedent and the anaphor as identity (see Dalrymple 1993, 2001). The representative entry in (29) is for the first person singular form *magam* 'myself'.

(29) *magam*: (↑PRED) = ‘PRO’
(↑PERS) = 1
(↑NUM) = SG
(↑CASE) = NOM
(↑PRON-TYPE)= REFL
(↑NUCL)= +
~ (SUBJ ↑)

The NUCL+ feature will require the reflexive to bind to a local antecedent (which cannot be a syntactic subject).

The entry *önmaga*$_1$ is a more or less run-of-the-mill reflexive. It agrees with its antecedent, it cannot occur as a subject, and it prefers bound variable readings.

(30) *önmagam*$_1$: (↑PRED) = ‘PRO’
(↑PERS) = 1
(↑NUM) = SG
(↑CASE) = NOM
(↑PRON-TYPE)= REFL
~ (SUBJ ↑)

The only important difference between (30) and (29) is that (30) lacks (or is underspecified for) the nuclear feature. This is to capture the fact that *önmaga*$_1$ is not necessarily subject to the Minimal Complete Nucleus binding condition (see 4.1).

What I have dubbed *önmaga*$_2$ is a special reflexive. It does not agree with its antecedent (or, rather, it is always third person singular), it can occur as a subject, and it prefers what looks like prima facie coreferential readings. It is this entry that I analyze here as a possessive reflexive.

What happens is that the extra nominal morphology (i.e., the prefix *ön-* ‘self’) reactivates the dormant possessive structure, which was lost during grammaticalization. This kind of special reanalysis is possible because, as we saw in 3.1., the reflexive stem has still retained the possessive morphology. The claim is that, in essence, *önmaga*$_2$ is analogous with the possessive expression *one’s self-representation*. The possessor is identified via the possessive agreement morphology, and the stem, *mag*, acts as some sort of a semantically bleached nominal.[7]

The proposed lexical entry is as follows, once again for the first person singular form:

[7] In (31) below, I assume a simplified f-structure analysis of the Hungarian possessive construction. See É. Kiss (2002) for a more detailed discussion of the data. Laczkó (2007) is a recent LFG-theoretic analysis of Hungarian possessives.

(31) *önmagam*$_2$: ($\uparrow$PRED) =‘SELF-REPRESENTATION$_i$’
($\uparrow$PERS) = 3
($\uparrow$NUM) = SG
($\uparrow$CASE) = NOM
($\uparrow$POSS PRED) = ‘PRO$_k$’
($\uparrow$POSS PERS) = 1
($\uparrow$POSS NUM) = SG

This analysis gives us an immediate account of the basic facts we observed. Since *önmagam*$_2$ is not a true anaphor, but a possessive structure, we expect it to be grammatical as a subject. What is more, we expect it to trigger constant third person singular agreement on the verb. (31) also describes the fact that this entry is not used in cases of semantic identity with the antecedent, but in ‘representations of the self’ contexts. Notice that the ‘antecedent’ now is referentially identified with the possessor buried inside the complex possessive structure of the reflexive. Thus, strictly speaking, what I described here somewhat sloppily as coreference between the reflexive and the antecedent is not direct coreference, but only a referential link between an individual and its representation via the underlying abstract possessive relation.

What the analysis does not capture is why the bound variable reading (between the antecedent and the possessor inside the structure of the reflexive) does not seem to be allowed in cases of VP-ellipsis. It may turn out that this really is just dispreference, contingent on the fact that this possessive structure arguably bears a level of idiomaticity.

Finally, let me add a note about to what extent the possessive analysis is motivated. I mentioned in the introduction that Beavers & Koontz-Garboden (2006) reject the possessive analysis of the English colloquial *his ass*, and treat it instead as a pronominal. They in fact entertain the idea of an analysis which would be analogous with (31) above, but then they reject it on the basis of the following minimal pair:

(32) a. *Mary had her office painted, and Jane had hers remodeled.*
b. **John got his ass a pedicure, and Pat got his a manicure*.

They argue that (32b) is ungrammatical because *his ass* is not a possessive structure. But the conclusion is not necessary, compare now (33a) and (33b):

(33) a. *My car is faster than John’s*.
b. **London’s fair city is nicer than Dublin’s*.

What I believe makes (33b) unacceptable is the general drive to avoid breaking up the internal structure of idiomatic units. *Dublin’s fair city* clearly does not encode a true possessive relation, which makes this noun phrase somewhat idiomatic. But it still is a possessive construction formally. I assume that similar considerations apply to the proposed entry for *önmaga*$_2$.

7. Summary and outlook

I started this paper by pointing out that Faltz's (1985) typology does not seem to cover reflexives that are the more complex versions of highly grammaticalized body part reflexives functioning as primary reflexive strategies in their respective languages. Whereas the prime reflexive is nuclear in accordance with the typology, its more complex version need not necessarily be nuclear.

On the basis of the analysis of the Hungarian body part reflexive *maga* and its more complex counterpart *önmaga*, I argued that what happens is that the extra nominal morphology (the prefix *ön*- 'self') reactivates the underlying possessive structure, and creates a special reflexive form. A similar analysis is proposed in Kornfilt (2001) for the Turkish *kendisi*, and possibly this analysis can also be extended to the Chinese *ta ziji*.

REFLEXIVES	PRONOMINAL	COMPLEX	POSSESSIVE
NORWEGIAN	*seg*	*seg selv*	
ENGLISH		*himself*	
HUNGARIAN		*maga*	*önmaga*
TURKISH		*kendi*	*kendisi*
CHINESE		*ziji*	*ta-ziji*

Table 2.

Table 2 gives an overview of the results. In essence, reflexives in the *possessive* column fall outside of Faltz's (1985) typology, and they add a new dimension to it.

I argued furthermore that when the possessive structure is triggered on *önmaga*, then we trigger at the same time a reading which targets representations of the self, rather than asserting identity with the self. It remains to be seen to what extent this property carries over to other reflexives with an active possessive structure. It is interesting to note nevertheless that languages that do not have possessive reflexives employ primary complex reflexives in 'representations of the self' contexts, as has been shown, among others, for the English *himself* by Jackendoff (1992), for the Dutch *zichzelf* by Reuland (2001), and for the Norwegian *seg selv* by Lødrup (2007). In contrast, the primary Hungarian complex reflexive, *maga*, does not allow for such readings. This suggests that one driving force behind the maintained interest in employing possessive reflexives is the need to have a form specialized for encoding dependencies which do not involve complete semantic identity between the antecedent and the reflexive.

Acknowledgements

I am grateful to Huba Bartos, Martin Everaert, Volker Gast, Ekkehard König, Tibor Laczkó and the participants of LFG09 in Cambridge for comments on various aspects of this proposal. Special thanks are due to the editors of this volume for their comments on the manuscript. Any remaining erros are mine. This research was supported by the 72983 OTKA grant on the Lexical-Functional Grammar approach to the Hungarian language.

Bibliography

Beavers, John & Koontz-Garboden, Andrew. 2006. A universal pronoun in English? *Linguistic Inquiry* 37 (3). 503-513.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Büring, Daniel. 2005. *Binding Theory*. Cambridge: CUP.

Cole, Peter; Hermon, Gabriella & Huang, C.-T. James (eds.) 2001. *Long-Distance Reflexives. Syntax and Semantics 33*. San Diego: Academic Press.

Cole, Peter; Hermon, Gabriella & Lee, Cher Leng. 2001. Grammatical and discourse conditions on long distance reflexives in two Chinese dialects. In P. Cole et al. (eds.) 2001. 1-46.

Culicover, Peter W. & Jackendoff, Ray. 2005. *Simpler Syntax*. Oxford: OUP.

Dalrymple, Mary. 1993. *The Syntax of Anaphoric Binding*. Stanford: CSLI.

Dalrymple, Mary. 2001. *Lexical Functional Grammar*. San Diego: Academic Press.

Faltz, Leonard. 1985. *Reflexivisation: A study in Universal Syntax*. New York: Garland.

É. Kiss, Katalin. 1994. Sentence structure and word order. In F. Kiefer & K. É. Kiss (eds.) *The syntactic Structure of Hungarian. Syntax and Semantics 27*. San Diego: Academic Press. 1-90.

É. Kiss, Katalin 2002. *The Syntax of Hungarian*. Cambridge: CUP.

Enç, Mürvet. 1989. Pronouns, licensing, and binding. *Natural Language and Linguistic Theory* 7 (1). 51-92.

Evans, Gareth. 1980. Pronoun*s*. *Linguistic Inquiry* 11(2). 337-362.

Everaert, Martin & Szendrői, Kriszta. 2002. Hungarian reflexive anaphors. Talk at the Linguistic Institute of the Hungarian Academy of Sciences.

Jackendoff, Ray. 1992. Mme. Tussaud meets the Binding Theory. *Natural Language and Linguistic Theory* 10 (1). 1-31.

Kiparsky, Paul. 2002. Disjoint reference and the typology of pronouns. In I. Kaufmann & B. Stiebels (eds.) *More than Words*. Berlin: Akademie Verlag. 179-226.

Kornfilt, Jaklin. 2001. Local and Long-Distance Reflexives in Turkish. In P. Cole et al. (eds.) 197-226.

König, Ekkehard & Gast, Volker. 2006. Focused assertion of identity: a typology of intensifiers. *Linguistic Typology* 10 (2). 223-276.

Laczkó, Tibor. 2007. Revisiting possessors in Hungarian DPs: A new perspective. In: M. Butt and T.H. King (eds.) *Proceedings of the LFG07 Conference*. Stanford: CSLI. 343-362. http://csli-publications.stanford.edu/LFG/12/lfg07.html

Lødrup, Helge. 2007. A new account of simple and complex reflexives in Norwegian. *The Journal of Comparative Germanic Linguistics* 10 (3). 183-201.

Pan, Haihua. 1998. Closeness, prominence, and binding theory. *Natural Language and Linguistic Theory* 16 (4). 771-815.

Pica, Pierre. 1987. On the nature of the reflexivization cycle. In J. McDonough & B. Plunkett (eds.) *Proceedings of the Seventeenth Annual Meeting of the North Eastern Linguistic Society*. *Volume 17*. 483-500.

Rákosi, György. 2006. *Dative Experiencer Predicates in Hungarian*. PhD. Dissertation. Utrecht. Published as volume 146 of the LOT Dissertation Series. http://www.lotpublications.nl/publish/issues/Rakosi/index.html

Reinhart, Tanya. 1983. *Anaphora and Semantic Interpretation*. London & Sydney: Croom Helm.

Reinhart, Tanya. 2006. *Interface Strategies. Optimal and Costly Computations. Linguistic Inquiry Monographs* 45. Cambridge, MA: The MIT Press.

Reuland, Eric. 2001. Primitives of Binding. *Linguistic Inquiry* 32 (3). 439-492.

AUTOMATIC ACQUISITION OF LFG RESOURCES FOR GERMAN - AS GOOD AS IT GETS

Ines Rehbein and Josef van Genabith
Universität des Saarlandes Dublin City University

Proceedings of the LFG09 Conference

Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

We present data-driven methods for the acquisition of LFG resources from two German treebanks. We discuss problems specific to semi-free word order languages as well as problems arising from the data structures determined by the design of the different treebanks. We compare two ways of encoding semi-free word order, as done in the two German treebanks, and argue that the design of the TiGer treebank is more adequate for the acquisition of LFG resources. Furthermore, we describe an architecture for LFG grammar acquisition for German, based on the two German treebanks, and compare our results with a hand-crafted German LFG grammar.

1 Introduction

Traditionally, deep, wide-coverage linguistic resources are hand-crafted and their creation is time-consuming and costly. Much effort has been made to overcome this problem by automatically inducing linguistic resources like rich, deep grammars, lexicons and subcategorisation frames from corpora. Most work so far has concentrated on English, like that of Hockenmaier and Steedman [2002], Nakanishi et al. [2004] and Cahill et al. [2002, 2004]. They present successful approaches for the acquisition of deep linguistic resources from the Penn-II treebank, using different grammar frameworks like CCG, HPSG and LFG. English, however, is a configurational language, where strict word-order constraints help to disambiguate predicate-argument structure. Porting these approaches to a semi-free word order language, we have to ask: How good can it get? Can we expect similar results when dealing with (semi-) free word order? Can data-driven methods cope when dealing with ambiguous data structures and sparse data, caused by a rich(er) morphology in combination with case syncretism? And, furthermore, what impact does treebank design have on the automatic acquisition of linguistic resources like deep grammars?

This paper describes approaches to treebank-based acquisition of LFG resources for a semi-free word order language, based on the method of Cahill et al. [2002, 2004, 2008], Burke et al. [2004] and O'Donovan et al. [2005], who presented the large-scale acquisition of LFG grammars and lexical resources from the English Penn-II and Penn-III treebanks. They also presented work on data-driven multilingual unification grammar development for Spanish, Chinese and German. While results point to treebank-based grammar acquisition being a universal method, results for other languages are by far lower than the ones achieved for English and the English Penn treebank.

There are different possible reasons for this: first of all, the size of the English Penn-II treebank, which is much larger than most treebanks for other languages, might be responsible for the good results on English. Another reason might be the configurational English word order, where strict constraints determine the grammatical function of a lexical unit in a certain surface position. Finally, the good results for English might be due to the data structures employed in the Penn-II treebank, which might be optimised for the task at hand and thus improve performance on the English data.

In this paper we develop different f-structure Annotation Algorithms for German, based on two German treebanks with crucially different annotation schemes, adapted to feature sets of varying granularity as represented in three different gold standards. We discuss problems specific to the annotation schemes of the two treebanks as well as to language-specific properties of German, where the variability in word order and the richer morphology (compared to English) often result in data sparseness, causing severe problems for data-driven methods. Finally, we compare the performance of our data-driven grammar acquisition architectures with the hand-crafted German ParGram LFG of Dipper [2003], Rohrer and Forst [2006], and Forst [2007].

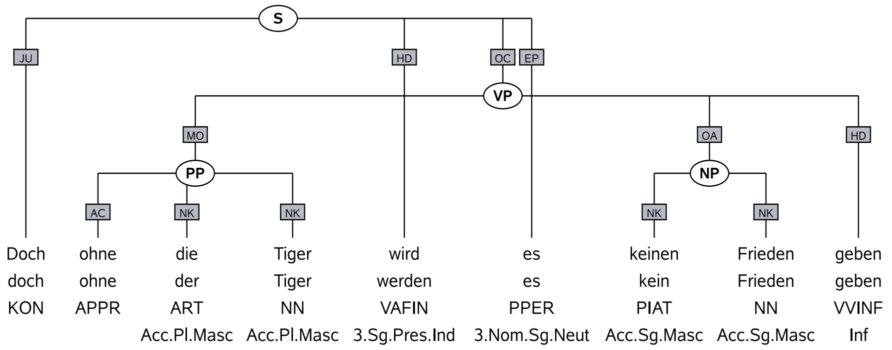
The paper is structured as follows: Section 2 gives an overview of typological properties of German and their representation in two different German treebanks. Section 3 describes the LFG grammar acquisition architecture for German, focusing on the differences to the work of Cahill et al. [2003, 2005] and Cahill [2004]. Section 4 reports on the automatic generation of LFG f-structures and discusses problems specific to semi-free word order and to the design of the German treebanks. Section 5 presents a comparison of our best automatically acquired LFG grammar with related work, namely the hand-crafted ParGram LFG for German. The last section concludes.

2 Typological Properties of German and their Representation in Two German Treebanks

German, like English, belongs to the Germanic language family. Despite being closely related, there are crucial differences between the two languages. One of them is the semi-free word order in German, which contrasts with the more configurational English; another, but related difference concerns the richer morphology in German, compared to the rather impoverished English morphology. Both properties are reflected in the treebank data structures used to represent syntactic analyses of the particular languages.

2.1 TiGer and TüBa-D/Z: Two German Treebanks

The TiGer treebank [Brants et al., 2002] and the TüBa-D/Z [Telljohann et al., 2005] are two German treebanks with text from the same domain, namely newspaper text. Both treebanks are annotated with phrase structure trees, dependency (grammatical relation) information and POS tags, using the Stuttgart Tübingen Tag Set (STTS) [Schiller et al., 1995]. Differences regard the set of categorial node labels used for syntactic annotation and the set of grammatical function labels. TiGer annotates 25 different syntactic categories and distinguishes between 44 different grammatical functions, while the TüBa-D/Z uses 26 different syntactic categories and 40



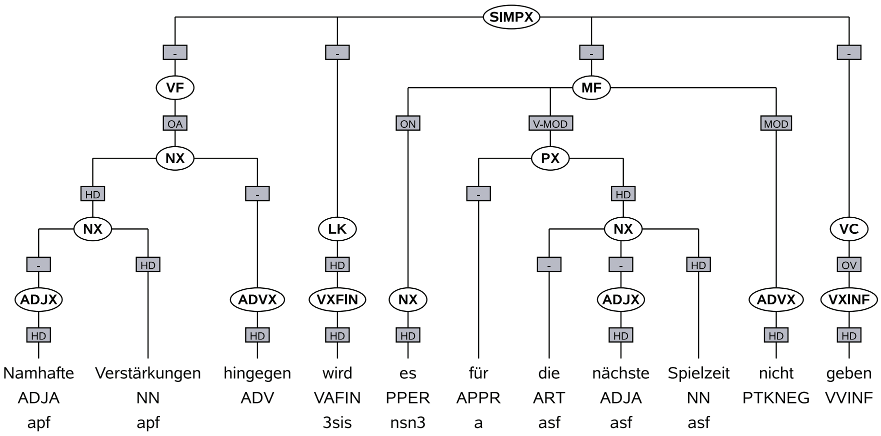
"But without the Tigers there will be no peace"

Figure 1: TiGer treebank tree

grammatical function labels. The main differences between the two treebanks are: (1) the flatter annotation in TiGer compared to the more hierarchical annotation in TüBa-D/Z, (2) the annotation of unary nodes in the TüBa-D/Z and no unary nodes in TiGer, (3) TüBa-D/Z uses topological fields to annotate the semi-free German word order, which allows for three possible sentence configurations (verb-first, verb-second and verb-final), and (4) TiGer annotates Long Distance Dependencies through crossing branches, while TüBa-D/Z encodes LDDs with the help of grammatical function labels (see Figures 1 and 2).

3 Automatic Annotation of LFG F-Structures

Cahill et al. [2003, 2004, 2005, 2008] presented a modular architecture for automatically annotating the English Penn-II treebank with LFG f-structures (Figure 3), which enables them to automatically extract deep, wide-coverage grammars which yield results in the same range as the best hand-crafted grammars for English [Briscoe and Carroll, 2002, Kaplan et al., 2004]. The f-structure Annotation Algorithm (AA) exploits lexical head information, and categorial, configurational and functional information as well as traces and co-indexation annotated in the Penn-II treebank. After determining the head of each constituent, the main module of the AA uses *left-right context annotation principles* to assign the most probable f-structure equation to each node in the tree (Figure 3). These principles express annotation generalisations and have been hand-crafted by looking at the most frequent grammar rules for each node in the Penn-II treebank and are also applied to unseen low-frequency rules. A sample partial left-right context annotation rule for NPs is given in Table 1. The left-context rule states that all adjectives or adjectival phrases to the left of the head of an NP should be annotated as an adjunct, while the right-context rule specifies that an NP to the right of the head of an NP is an



"However, there won't be considerable reinforcements for the next playing season."

Figure 2: TüBa-D/Z treebank tree

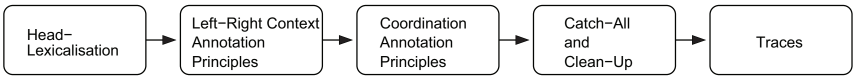


Figure 3: Architecture of the English f-structure Annotation Algorithm (AA)

apposition. The creation of these left-right-context rules needs linguistic expertise and crucially depends on configurational properties of English.

left-context	head	right-context
JJ, ADJP: $\downarrow = \in \uparrow$ ADJUNCT	NN, NNS, ... $\uparrow=\downarrow$	NP: $\downarrow = \in \uparrow$ APP

Table 1: Left-right context annotation rule used in the English AA

Coordinations are treated seperately. After adding f-structure equations to all nodes in the tree, the *Catch-All and Clean-Up* module deals with overgeneralisations. Finally, traces are resolved.

The German LFG AA, like the English one, is highly modularised and proceeds as follows (Figure 4). First it reads in the treebank trees encoded in the NEGRA export format and converts each tree into a tree object. Then it applies head-finding rules which we developed in the style of Magerman [1995], in order to determine the head of each local node.[1] The head-finding rules specify a set of candidate heads, depending on the syntactic category of the node, and also the

[1] TiGer provides head annotation for all categorial nodes except NPs, PPs and PNs. Due to the flat annotation in TiGer, partly resulting from the decision not to annotate unary nodes, the problem of identifying the correct head for those nodes is more severe than for the TüBa-D/Z, where the more hierarchical structure results in smaller constituents which, in addition, are all head-marked. When annotating original treebank trees, the head-finding rules are applied to NP, PP and PN nodes; when

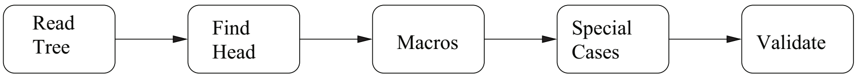


Figure 4: Architecture of the German f-structure Annotation Algorithm

direction (left/right) in which the search should proceed. For prepositional phrases, for example, we start from the left and look at all child nodes of the PP. If the left-most child node of the PP has the label KOKOM (comparative particle), we assign it as the head of the PP. If not, we check if it is a preposition (APPR), a preposition merged with a determiner (APPRART), an apposition (APPO), and so on. If the left-most child node does not carry one of the candidate labels, we take a look at the next child node, working our way from left to right.

For some of the nodes these head-finding rules work quite well, while for others we have to accept a certain amount of noise. This is especially true for the flat NPs in the TiGer treebank. A *Special Cases* module checks these nodes at a later stage in the annotation process and corrects possible errors made in the annotation.

After determining the heads, the tree is handed over to the *Macros* module which assigns f-structure equations to each node. This is done with the help of macros. Sometimes these macros overgeneralise and assign an incorrect grammatical function. In order to deal with this, the *Special Cases* module corrects inappropriate annotations made by the *Macros* module. Finally the *Validation* module takes a final look at the annotated trees and makes sure that every node has been assigned a head and that there is no node with two child nodes carrying the same governable grammatical function.

The most important difference in the design of the English and the German AAs concerns the application of left-right context annotation rules described above. For English, these rules successfully specify the correct annotation for the majority of local nodes in a given tree. For German, however, these rules do not work as well as for English. Table 2 illustrates this point by showing different possibilities for the surface realisation of a (rather short) German sentence. Some of the examples are highly marked, but all of them are possible surface realisations of (1).

(1) Die Anklage legt ihm deshalb Betrug zur Last.
the prosecution lies him therefore fraud to the burden.
The prosecution therefore charges him with fraud.

The f-structure-annotated grammar rule for the sentence in (1) (Figure 5) tells us that the first NP *Die Anklage* (the prosecution) is the subject of the sentence,

running the AA on parser output trees with erroneous or no GF labels in the trees, we also make use of head-finding rules for other syntactic categories.

In TüBa-D/Z, heads are marked for most categorial nodes. However, there are some open issues, like the one concerning the head of the middle field or of proper name nodes, or the annotation of appositions, which are considered to be referentially identical and therefore bear no head marking in the TüBa-D/Z.

$$\begin{array}{llllllll} \mathrm{S} & \rightarrow & \mathrm{NP} & \mathrm{VVFIN} & \mathrm{PPER} & \mathrm{PROAV} & \mathrm{NN} & \mathrm{PP} \\ & & \uparrow \mathrm{SUBJ}{=}\downarrow & \uparrow{=}\downarrow & \uparrow \mathrm{DA}{=}\downarrow & \downarrow{\in}\uparrow \mathrm{MO} & \uparrow \mathrm{OA}{=}\downarrow & \uparrow \mathrm{OP}{=}\downarrow \end{array}$$

Figure 5: Grammar rule and f-structure equations for the sentence in (1)

Die Anklage	legt	ihm	deshalb	Betrug	zur Last.
Betrug	legt	ihm	deshalb	die Anklage	zur Last.
Ihm	zur Last	legt	die Anklage	deshalb	Betrug.
Zur Last	legt	ihm	die Anklage	deshalb	Betrug.
Deshalb	legt	ihm	die Anklage	Betrug	zur Last.
...	...	...	...	...	...

Table 2: Variable word order in German (sentence (1))

while the noun *Betrug* (fraud) should be annotated as an accusative object, and the pronominal adverb *deshalb* (therefore) is an element of the modifier set. Table 2, however, illustrates that these constituents can occur in very different positions to the left or right of the head of the sentence. This shows that, unlike for a strongly configurational language such as English, the specification of left-right-context rules for German is not very helpful.

Instead of developing horizontal and strongly configurational context rules, the AA for German makes extended use of macros, using different combinations of information such as part-of-speech (POS) tags, node labels, edge labels and parent node labels (as encoded in the TiGer and TüBa-D/Z treebanks). First we apply more general macros assigning functional annotations to each POS, syntactic category or edge label in the tree. More specific macros, such as the combination of a POS tag with the syntactic node label of the parent node or a categorial node with a specific grammatical function label, can overwrite these general macros. The order of these macros is crucial, dealing with more and more specific information. Some of the macros overwrite information assigned before, while others only add more information to the functional annotation.

To give an example, consider the POS tag ART (determiner). The first macro is triggered by this POS tag and assigns the f-structure equation $\uparrow=\downarrow, \downarrow \textit{det-type} = def$. The next macro looks at combinations of POS tags and grammatical function (GF) labels and, for a determiner with the label NK (noun kernel), adds the equation $\uparrow \textit{spec} : \textit{det} =\downarrow$, while the same POS tag gets assigned the functional equation $\downarrow\in\uparrow \textit{spec} : \textit{number}$ when occurring with the edge label NMC (numerical component). The annotation for the combination of POS and grammatical function label can be overwritten when a more specific macro applies, e.g. one which also considers the parent node for a particular POS-GF-combination.

The determiner with edge label NK has so far been annotated with $headword, \downarrow \textit{det-type} = def, \uparrow \textit{spec} : \textit{det} =\downarrow$. This is overwritten with the f-structure equation $\uparrow obj : \textit{spec} : \textit{det} =\downarrow$, if it is the child of a PP node. This is due to the fact that the annotation guidelines of the TiGer treebank analyse prepositions as the head of a PP, while the head noun (and its dependents) inside the PP is annotated as the object of the preposition. Due to the flat annotation in the TiGer treebank, it is not helpful to use vertical context above the parent node level. The AA makes heavy use of the *Special Cases* module, where further annotation rules are specified for most syntactic categories. One tricky case is that of NPs, which have a totally flat structure in the TiGer treebank. There are many cases where the information about POS tag and grammatical function label is not sufficient, and neither is their relative position to the head of the phrase. In those cases the presence or absence of other nodes decides the grammatical function of the node in question.

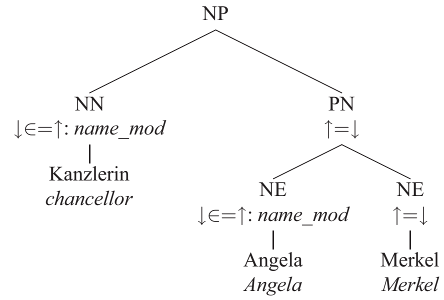


Figure 6: NP-internal structure in TiGer (PN=head)

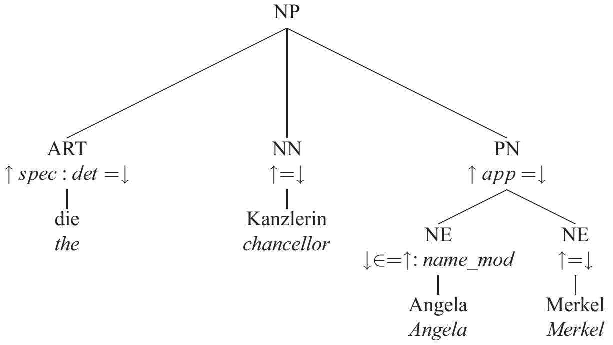


Figure 7: NP-internal structure in TiGer (PN=apposition)

To illustrate this, consider the three examples in Figures 6-8. All three examples show an NP with a noun child node followed by a proper name (PN) node, but where the grammatical annotations differ crucially. In Figure 6, the PN is the head of the NP. In Figure 7, where we have a determiner to the left of the noun (NN), the noun itself is the head of the NP, while the PN is an apposition. The third example (Figure 8) looks pretty much like the second one, with the exception that *Merkel* is in the genitive case. Here the PN should be annotated as a genitive attribute. This is not so much a problem for the annotation of the original treebank trees where we have both the correct grammatical function labels as well as morphological information. For parser output, however, morphological information is not available and the grammatical functions assigned are often incorrect. In Section 4.2.1

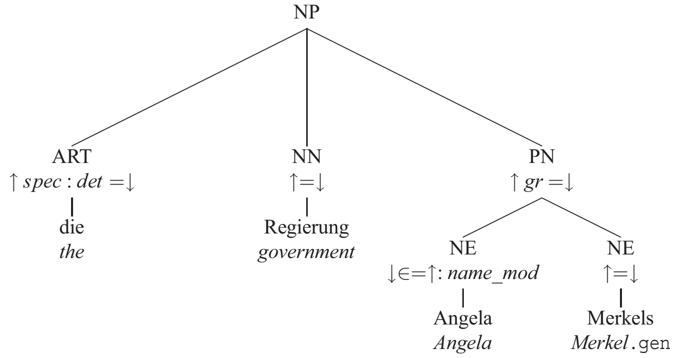


Figure 8: NP-internal structure in TiGer (PN=genitive to the right)

we will return to this issue und discuss the reason for the missing morphological information in the parser output.

3.1 Differences between our AA for German and Preliminary Work

The annotation algorithm for German presented in this chapter is based on and substantially revises and extends preliminary work by Cahill et al. [2003, 2005] and Cahill [2004]. The AA by Cahill et al. provides annotations for a rather limited set of grammatical functions only (26 grammatical functions: 11 governable functions, 10 non-governable functions and 5 atomic features). We created a new gold standard f-structure bank containing 250 sentences from the TiGer treebank, the TIGER250, which uses a substantially extended set of grammatical functions and features (46 grammatical functions: 14 governable grammatical functions, 13 non-governable grammatical functions and 19 atomic features). As a result, the annotated resources contain richer linguistic information and are of higher quality and usefulness compared to the one of Cahill et al. [2003, 2005] and Cahill [2004]. Our annotation algorithm also makes use of a valency dictionary in order to distinguish between stative passive constructions and the German Perfekt with *sein* ’to be’.

We also adapted the AA to the feature set used in the TiGer DB[2] [Forst et al., 2004] (Dependency Bank) and a hand-crafted gold standard from the TüBa-D/Z[3] (TUBA100).

[2] The TiGer DB distinguishes 52 different grammatical features. We use a slightly modified version without the distinction between different prepositional objects, and without morphological features or compound analysis.

[3] The TüBa-D/Z gold standard was semi-automatically created by Heike Zinsmeister and Yannick Versley, using the conversion method of Versley [2005] on 100 randomly selected trees from the TüBa-D/Z. The feature set is similar to the TiGer DB.

4 LFG F-Structure Annotation and Evaluation on Two German Treebanks

For German, we adapted the AA to the node and edge labels of the two German treebanks. As described above, word order variation in German does not allow to make strong use of configurational information as in the English AA. Instead, we heavily rely on the grammatical function labels in the trees. This works well when annotating original treebank trees, but causes many problems when applied to parser output. State-of-the-art parsing results as presented in the PaGe Shared Task on Parsing German [Kübler, 2008] are in the range of 58-70% F-score for TiGer and 75-84% for TüBa-D/Z.[4] The differences in annotation schemes do not allow for a direct comparison of parsing results, but the message is clear: for both treebanks automatically assigned syntactic nodes and, even more important, grammatical function labels are to a great extent error-prone, which defines an upper bound for treebank-based parsing into f-structures using the automatic annotation algorithm.

Section 4.2 presents parsing experiments with automatic LFG f-structure annotation based on TiGer and TüBa-D/Z, and evaluates the generated f-structures against hand-crafted gold standards from the TiGer treebank (TiGer DB, TIGER250) and from the TüBa-D/Z (TUBA100). However, before applying the AA to parser output we want to test its performance on gold standard syntax trees.

4.1 Results for LFG F-Structure Annotation on Gold Standard Syntax Trees

Table 3 shows results for automatic f-structure annotation on gold treebank trees for the sentences in the TiGer DB, the TIGER250 and the TUBA100.[5] Results for the TIGER250 and the TUBA100 are quite good, while results for the TiGer DB are around 10% lower. This is due to mapping problems between the TiGer DB and TiGer treebank. The sentences in the TiGer DB have been converted semi-automatically into a dependency-based triple format, using a large, hand-crafted LFG grammar for German [Dipper, 2003] and then manually corrected. The TiGer DB provides a very fine-grained description of linguistic phenomena in German,

	Prec.	**Rec.**	**F-Score**
TiGerDB	87.8	84.8	86.3
TIGER250	96.8	97.5	97.1
TUBA100	95.5	94.6	95.0

Table 3: Results for automatic f-structure annotation on gold treebank trees

[4] Results report constituent-based evalb labelled F-scores on syntactic nodes and grammatical function labels when using gold POS tags with gold GF labels as parser input

[5] We split the gold standards into development and test set, with 500 test set trees for the TiGer DB and 125 test trees for the TIGER250. Due to its limited size, we did not split the TUBA100.

but includes additional information which is not annotated in the TiGer treebank and thus cannot be derived automatically. This means that the TiGer DB-based evaluation is biased in favour of the hand-crafted LFG grammar of Dipper [2003].

4.2 Parsing German with Automatically Acquired LFG Grammars

In our experiments we use the Berkeley parser [Petrov and Klein, 2008], a language-agnostic parser which automatically refines and re-annotates the training data by applying split-and-merge operations, so that the likelihood of the transformed treebank is maximised. The Berkeley parser achieved the best results in the Shared Task on Parsing German (ACL 2008).

We removed the gold standard sentences from the treebanks and extracted two training sets with 25,000 sentences each. For TiGer we persued two different ways of resolving crossing branches in the trees: (1) by attaching the non-head child nodes higher up in the tree, following Kübler [2005], and (2) by splitting discontinuous nodes into smaller "partial nodes" [Boyd, 2007], a strategy which aims at preserving local tree structure while allowing the system to recover the original dependencies after parsing. With regard to GF labels we tested two different settings: in the first setting (Atomic) we merged categorial node labels with grammatical function labels and trained the parser on the new atomic labels. In the second setting (FunTag) we removed GF labels from the training data and trained the parser on syntactic categories only. The GF labels were then assigned in a post-processing step, using the SVM-based grammatical function labelling software by Chrupała et al. [2007]. We parsed the different test sets with the extracted grammars and, for the grammars without grammatical functions, let FunTag assign GFs to the parser output. The trees with grammatical function labels were passed over to the AA, where all nodes in the parse trees were annotated with LFG functional equations. Next we collected the equations and handed them over to a constrainst solver, which generated LFG f-structures.

4.2.1 Results

Table 4 shows constituent-based parsing results for the different test sets and settings (Atomic, FunTag) as well as results for f-structure evaluation. For the first setting, where we let the Berkeley parser assign the grammatical functions (Atomic), the two TiGer test sets yield constituent-based parsing results in the range of 76-79% (labelled F-score on syntactic categories) and 67-70% (including GF labels). Results for the TüBa-D/Z are more than 10% higher, which is an artifact of the different treebank annotation schemes and does not reflect parser output quality, as can be seen in the f-structure evaluation. On the f-structure level precision is in the range of 73-81%, while recall for the TüBa-D/Z f-structures is dramatically lower at around 45%. For the TiGer, we achieve a recall of 73.7% for TiGer DB and of 79.7% for the TIGER250 test set.

Parsing results for the Berkeley parser trained on TiGer syntactic nodes only

Constituent-based evaluation						
	Atomic			**FunTag**		
length<= 40	**F-score**	**F-score** GF	**POS acc.**	**F-score**	**F-score** GF	**POS acc.**
TiGerDB	79.3	70.2	96.0	81.0	70.9	97.0
TIGER250	76.6	66.9	95.4	79.3	68.4	96.5
TUBA100	89.3	80.2	96.5	89.2	76.3	96.4
f-structure evaluation						
	Atomic			**FunTag**		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
TiBerDB	73.0	73.9	73.4	76.1	65.1	70.2
TIGER250	81.4	79.7	80.5	87.6	67.5	76.3
TUBA100	76.9	45.1	56.9	75.8	39.3	51.7

Table 4: C-structure parsing results (labelled F-score without and with GF) and f-structure evaluation

(FunTag) are higher than for the atomic labels. For TüBa-D/Z, however, we observe better results when training on both syntactic categories and grammatical functions. The FunTag-assigned GFs yield better `evalb` results and a higher precision for the TiGer f-structures. For the TüBa-D/Z, precision is slightly lower than for f-structures generated from parser output where the Berkeley parser did the function labelling. The better precision for the TiGer f-structures comes at the cost of a decrease in recall. For the TüBa-D/Z f-structures, recall is even lower than before.

There are several reasons for the low recall for the TüBa-D/Z: (1) Due to its limited size the TUBA100 does not cover all relevant grammatical phenomena and therefore is not sufficient as a test set for grammar development, which is reflected in the low recall score. (2) Phrases without a clear dependency relation to the other constituents in the tree are attached directly to the root node in the TüBa-D/Z. The resulting tree structure makes it impossible for the AA to disambiguate the sentence and find a suitable dependency relation for the highly attached node, which means that these nodes are not represented in the f-structure, further lowering recall for the TüBa-D/Z. (3) NP internal structure in the TüBa-D/Z contains less information than in TiGer, where grammatical function labels distinguish genitive attributes, dative attributes and comparative complements. The missing information can be partly retrieved from morphological annotation, but this would require an extensive treebank transformation to make this information available to the parser. The grammars extracted from the treebanks do not include morphological information, which means that the TiGer grammars encodes more specific functional information than the TüBa-D/Z grammars.

Yet another reason for the lower recall for TüBa-D/Z f-structures can be found in the design of the grammatical function labels used in the annotation. While the original treebanks use roughly the same number of grammatical functions (44 in TiGer versus 40 in TüBa-D/Z; Table 5), some of the grammatical functions in the TüBa-D/Z occur only with a very low frequency. When comparing two smaller subsets of 2,000 gold treebank trees, we still find 42 of the 44 GFs in

	Gold all	**Gold 2000**	**Atomic**	**FunTag**
TiGer	44	42	41	40
TüBa-D/Z	40	33	31	19

Table 5: Number of different grammatical functions in TiGer/TüBa-D/Z gold trees and reproduced in the different parsing settings (Atomic/FunTag)

the TiGer set, while the TüBa-D/Z subset uses only 33 of the 40 GFs. For parser output the problem gets even worse. In the TiGer-trained parser output for the same subset of 2,000 sentences we find 41 different GF labels when the Berkeley parser assigns the grammatical functions, and 40 when FunTag does the GF labelling, while in a data set of the same size from the TüBa-D/Z, only 31 different GF labels are used in the parser output (Atomic), and the FunTag approach yields only 19 different grammatical functions. This leads to a crucial difference between the type of information encoded in the GF labels for the two treebanks: while TiGer labels describe the grammatical function of one node, in TüBa-D/Z the GF labels (besides the main grammatical functions such as subject and acusative or dative object) express dependency relations between different nodes in the tree, which are often positioned in different topological fields. As pointed out, some of the grammatical functions in the TüBa-D/Z occur with a very low frequency.[6] This poses a problem for machine learning methods, which rely on a sufficiently large set of training instances in order to achieve good performance on unseen data.

GF	**Atomic**	**FunTag**	**Atomic**	**FunTag**
	TiGer (2,000 sent.)		*TüBa-D/Z (2,000 sent.)*	
DA	52.5	74.9	56.8	27.2
OA	79.5	85.5	69.0	46.4
SB	90.0	88.4	85.2	72.1
ALL GF	93.1	94.4	91.9	88.3

Table 6: Evaluation of main grammatical functions in TiGer and TüBa-D/Z (dative object: DA/OD, accusative object: OA, subject: SB/ON)

Next we compare results for the main grammatical functions (subject, accusative and dative object) on 2,000 sentence test sets from TiGer and TüBa-D/Z (Table 6). For parser-assigned GFs, we observe better results for dative objects (DA/OD) for the parsing model trained on the TüBa-D/Z, while for subjects and accusative objects the TiGer-trained parser yields better results. The SVM-based FunTag shows poor performance on the TüBa-D/Z data, while for TiGer the function labeller outperformes the setting where the Berkeley parser does the GF assignment (Atomic). This divergent behaviour might be due to the different data

[6] OA-MODK (conjunct of modifier of accusative object), ON-MODK (conjunct of modifier of nominative object) and OADVPK (conjunct of modifier of ADVP object) occur only once in 27,125 sentences in TüBa-D/Z Release 3, OG-MOD (modifier of genitive object) 7 times, OADJP-MO (modifier of ADJP object) 8 times, OADVP-MO (modifier of ADVP object) 10 times, and FOPPK (facultative object of PP object) 17 times.

structures in the treebanks. The split into topological fields in the TüBa-D/Z takes away necessary context information, which is encoded in the feature set for the flat TiGer trees.

4.3 Different Approaches to Discontinuity and their Impact on F-Structure Annotation

Boyd [2007] presents an improved method for converting the crossing branches in TiGer into context-free representations by splitting up discontinuous nodes into marked "partial" nodes. She shows that the improved conversion results in more consistent trees and improves results in a labelled dependency evaluation for accusative, dative and prepositional objects. In her experiments, Boyd used an unlexicalised PCFG parsing model (LoPar, Schmid [2000]) with gold POS tags as parser input.

We applied the split-node conversion method to the TiGer data and trained the Berkeley parser on the converted training sets. Table 7 shows parsing results for the two conversion methods: (1) raised nodes and (2) split nodes. For the TiGer DB test set, results for the split-node conversion are slightly worse, while for the TIGER250 test set there is a small improvement of 1% F-score. For both data sets, however, the number of valid f-structures decreases considerably.

	Precision	**Recall**	**F-score**	**valid F-struc.**
		TiGer DB		
raised	73.0	73.9	73.4	82.4
split	71.8	72.0	71.9	71.0
		TIGER250		
raised	81.5	80.9	81.2	88.0
split	82.7	81.8	82.2	84.0

Table 7: f-structure evaluation on converted TiGer trees (raised- vs. split-node)

Boyd's split-node conversion works well for pure PCFG parsers like LoPar. The Berkeley parser, however, makes use of horizontal markovisation, which breaks up the original grammar rules and generates new rules which have not been seen in the training set. This also admits rules with only one of the two partial nodes, which means that a reconstruction of the original tree is impossible, and often leads to clashes during f-structure generation.

5 LFG Parsing: Related Work

This section discusses related work and shows how our research compares to the wide-coverage hand-crafted LFG grammar of Dipper [2003], Rohrer and Forst [2006], and Forst [2007] developed in the ParGram project [Butt et al., 2002]. The ParGram German LFG uses 274 LFG-style rules (with regular expression-based right-hand sides) and several lexicons with detailed subcategorisation information and a guessing mechanism for default lexical entries [Rohrer and Forst,

	ParGram			**TiGerDB**	**DCU250**
GF	**up. bound**	**log. lin.**	**low. bound**		
da	67	63	55	44	38
gr	88	84	79	71	87
mo	70	63	62	65	73
oa	78	75	65	69	63
quant	70	68	67	67	78
rc	74	62	59	34	30
sb	76	73	68	74	79
preds only	79.4	75.7	72.6	72.7	78.6
	coverage on the NEGRA treebank (>20,000 sentences)				
	81.5	81.5	81.5	88.2	88.7

Table 8: F-scores for selected grammatical functions for the ParGram LFG (upper bounds, log-linear disambiguation model, lower bounds) and for two automatically acquired TiGer grammars

2006]. Preprocessing in the experiments reported in Rohrer and Forst [2006] includes modules for tokenisation, morphological analysis and manual marking of named entities, before the actual parsing takes place. An additional disambiguation component based on maximum entropy models is used for reranking the output of the parser. Forst [2007] tested parser quality on 1,497 sentences from the TiGer DB and reported a lower bound, where a parse tree is chosen randomly from the parse forest, an upper bound, using the parse tree with the highest F-score (evaluated against the gold standard), as well as results for parse selection done by the log-linear disambiguation model.

Table 8 shows results for the ParGram LFG and for the automatically induced grammars on selected grammatical relations and on all grammatical functions excluding morphological and other features (preds only). The automatically induced TiGer DB and DCU250-style grammars were trained on the full TiGer treebank (>48,000 sentences, excluding the test data). We report results for the test sets from the TiGer DB and the DCU250.

The hand-crafted LFG outperforms the automatically acquired grammars on most GFs for the TiGer DB, but results are not directly comparable. The TiGer DB-based evaluation is biased in favour of the hand-crafted LFG. Named entities in the ParGram LFG input are marked up manually, while for our grammars these multiword units often are not recognised correctly and so are punished during evaluation, even if part of the unit is annotated correctly. Furthermore, the hand-crafted ParGram LFG grammar was used in the creation of the TiGer DB gold standard in the first place, ensuring compatibility as regards tokenisation and overall linguistic analysis.

F-scores for the DCU250 are in roughly the same range as the ones for the hand-crafted grammar. For high-frequency dependencies like subjects (sb) or modifiers (mo), results of the two grammars are comparable. For low-frequency depen-

	ParGram			**TiGerDB**	**DCU250**
GF	**up. bound**	**log. lin.**	**low. bound**		
da	67	63	55	58	50
gr	88	84	79	68	88
mo	70	63	62	63	77
oa	78	75	65	68	80
quant	70	68	67	58	69
rc	74	62	59	50	50
sb	76	73	68	76	85
preds only	79.4	75.7	72.6	76.0	84.4

Table 9: Precision for selected grammatical functions for the ParGram LFG and for the TiGer grammars

dencies like dative objects (da) or relative clauses (rc), however, the hand-crafted LFG outperforms the automatic LFG f-structure annotation algorithm by far. Coverage for the automatically acquired grammars is considerably higher than for the hand-crafted LFG grammar. Rohrer and Forst [2006] report a coverage of 81.5% (full parses) when parsing the NEGRA treebank, which contains newspaper text from the same newspaper as in the TiGer treebank. By contrast, the automatically acquired TiGer grammars achieve close to 90% coverage on the same data. On the TiGer treebank Rohrer and Forst [2006] report coverage of 86.4% full parses, raising the possibility that, as an effect of enhancing grammar coverage by systematically extracting development subsets from TiGer, the ParGram LFG is tailored closely to the TiGer treebank.

The DCU250 test set is equally biased towards the TiGer treebank-based LFG resources, as it only represents what is encoded (directly or implicitly) in the TiGer treebank. The truth is somewhere in between: The TiGer DB evaluation of the treebank-based LFG resources attempts to a limited extent to counter the bias of the original TiGer DB resource towards the hand-crafted LFG grammar by removing distinctions which cannot be learned from TiGer data only, and by relating TiGer DB to (some of) the original TiGer tokenisation using the version prepared by Boyd et al. [2007]. The resulting resource still favours the hand-crafted LFG resources, which outperform the treebank-based resources by about 3% points absolute. Looking at precision, results for the TiGer grammars are more or less in the same range as the F-scores for the Pargram LFG (Table 9).[7]

5.1 Discussion

Our automatically extracted grammars yield better coverage than the hand-crafted LFG of Dipper [2003], Rohrer and Forst [2006] and Forst [2007], but with regard to F-score the ParGram LFG still outperforms the automatically acquired gram-

[7] Unfortunately, Forst [2007] does not report results for precision and recall.

mars. The lower results for our grammars are not due to low precision: Table 9 contrasts F-scores for the Pargram LFG with results for precision as achieved by the automatically acquired TiGer grammars. Future work should therefore focus on improving recall in order to achieve results comparable with or better than hand-crafted grammars. One promising approach is the one of Seeker [2009], who describes a grammatical function labeller based on Integer Linear Programming (ILP). Seeker presents a two-step approach, consisting of a classification step and a selection step. During classification, the probability distribution over all possible labels for each node in the tree is computed, using a maximum entropy classifier. During selection, the overall probability of the whole tree is optimised, where the ILP-based approach allows the developer to implement hard constraints (e.g.: no more than one subject per local tree). First results show that global optimisation in combination with linguistically motivated constraints improves precision and coverage. F-scores for f-structure evaluation on the TiGer DB increase to more than 75%, while coverage was raised from around 88% to more than 96%.

An unsolved problem is the encoding of LDDs in treebank annotation schemes for (semi-) free word order languages. Currently, neither the TiGer treebank and even less so the TüBa-D/Z way of representing non-local dependencies can be learned successfully by statistical parsers. An approach to resolving LDDs at the f-structure level was described in Cahill et al. [2004] and Cahill [2004] and successfully implemented as part of the English treebank-based LFG acquisition and parsing architectures. However, the method of Cahill et al. relies on complete f-structures, which means that the recall problem must have been solved before we can reliably and profitably compute LDDs on f-structure level for German.

6 Conclusions

We presented two architectures for the automatic acquisiton of LFG resources, based on two German treebanks. Compared to a hand-crafted German LFG, our method yields higher coverage and comparable results for the high-frequency grammatical functions, while for the less frequent GFs the hand-crafted grammar clearly outperforms the automatic approach.

We have outlined a number of problems for treebank-based f-structure annotation for German: (1) The semi-free word order in German rules out the use of configurational information for f-structure annotation. (2) Parsing results for German, especially for GF assignment, are not reliable enough to support accurate f-structure annotation. (3) Our alternative approach to assign GF labels using an SVM-based function labeller achieves high precision, but at the cost of recall. This is due to missing context sensitivity of the function labeller, resulting in the assignment of conflicting GFs.

We showed that particular treebank encoding schemes have a strong impact on the usability of the resources. We argue that the GF label set in the TüBa-D/Z, which has been designed with the aim of expressing dependency relations between different nodes in the tree, is less adequate for the automatic acquisition of LFG resources than the label set in TiGer. The GF labels in the TüBa-D/Z are harder to learn and also encode less specific grammatical information than the ones in TiGer.

The task of automatically inducing linguistic resources from (semi-) free word order languages is much harder than for more configurational languages like English. Future research needs to address the problem of automatic GF assignment which for German is far more important than for configurational languages (one promising line of research has been outlined in Section 5.1). Only then can we expect to automatically induce high-quality linguistic resources for languages other than English and other configurational languages.

References

Adriane Boyd. Discontinuity revisited: An improved conversion to context-free representations. In *Proceedings of the Linguistic Annotation Workshop (LAW 2007)*, pages 41–44, Prague, Czech Republic, 2007.

Adriane Boyd, Markus Dickinson, and Detmar Meurers. On representing dependency relations – insights from converting the German TiGerDB. In *Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories (TLT-07)*, pages 31–42, Bergen, Norway, 2007.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The TIGER Treebank. In Erhard W. Hinrichs and Kiril Simov, editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, pages 24–42, Sozopol, Bulgaria, 2002.

Ted Briscoe and John Carroll. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-02)*, pages 1499–1504, Las Palmas, Canary Islands, 2002.

Michael Burke, Olivia Lam, Aoife Cahill, Rowena Chan, Ruth O'Donovan, Adams Bodomo, Josef van Genabith, and Andy Way. Treebank-based acquisition of a chinese lexical-functional grammar. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC-18)*, pages 161–172, Tokyo, Japan, 2004.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. The parallel grammar project. In *Proceedings of COLING-02 Workshop on Grammar Engineering and Evaluation*, pages 1–7, Taipei, Taiwan, 2002.

Aoife Cahill. *Parsing with Automatically Acquired, Wide-Coverage, Robust, Probabilistic LFG Approximations*. PhD dissertation, School of Computing, Dublin City University, Dublin, Ireland, 2004.