Incremental Identification of Inflectional Types

Petra Barg and **James Kilbury**
Heinrich-Heine-Universität Düsseldorf,
Seminar für Allgemeine Sprachwissenschaft,
Universitätsstr. 1, D-40225 Düsseldorf, Germany
E-mail:{barg,kilbury}@ling.uni-duesseldorf.de

Abstract

We present an approach to the incremental accrual of lexical information for unknown words that is constraint-based and compatible with standard unification-based grammars. Although the techniques are language-independent and can be applied to all kinds of information, in this paper we concentrate on the domain of German noun inflection. We show how morphological information, especially inflectional class, is successfully acquired using a type-based HPSG-like analysis. Furthermore, we sketch an alternative strategy which makes use of finite-state transducers.

1 Introduction

Systems for natural language processing must deal adequately with “unknown” words, i.e. lexemes that either have been newly coined or else have not been included in a particular lexicon (cf. Kilbury et al. (1994)). Rather than simply regarding unknown words as noise, our system instead uses their context as a source for the systematic accrual of lexical information that can then be utilized.

Our approach differs in significant respects from those of other investigators. It is designed for unification-based grammar formalisms with typed feature stuctures as in HPSG and is not restricted to simple morphosyntactic features. In contrast to statistical approaches like that of (Brent, 1991), which often do not work incrementally and are intended for the application to large corpora, ours instead aims at a detailed grammatical analysis of individual sentences with a maximal use of their information. While systems like that of (Hahn et al., 1996) deal with the general acquisition of concepts, we are concerned exclusively with the acquisition of structural linguistic information.

Although we deal here with German noun inflection, in a framework close to that of (Riehemann, 1998) and (Koenig, 1999), the techniques are language-independent and apply to other kinds of lexical information as well, as is shown in (Walther and Barg, 1998) with respect to valency information. Thus, in contrast to (Ehrlich and Rapaport, 1997), who employ tailored algorithms for the acquisition of information about nouns and verbs, we introduce an approach that is completely general with respect to the kind of structural linguistic information acquired.

2 German noun-inflection classes

There is a vast literature on German noun inflection represented in recent studies by (Cahill and Gazdar, 1999), (Clahsen, 1999), and (Neef, 1998). Here we summarize only essential points and ignore highly irregular and archaic inflections (cf. figure 4 below).

German nouns bear gender (masculine, feminine, neuter) and are inflected for number (singular, plural) and case (nominative, accusative, dative, genitive). With the exception of class NWN (e.g. masc *Bauer* ‘farmer’, with gen sg *Bauern*), all nonfeminine nouns build genitive singular with -*s*.

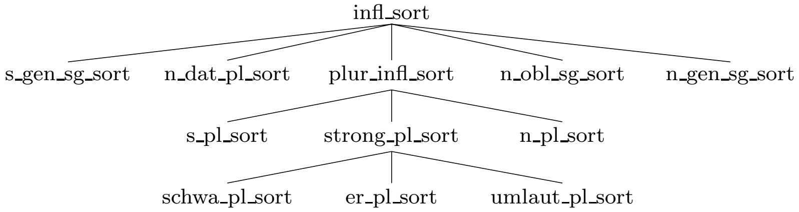
The “regular” (cf. Clahsen (1999)) but “atypical” (cf. Wunderlich (1999)) nouns of class NA (e.g. *Auto* ‘car’) build their plural forms in -*s*. The plural forms of all other (i.e. “typical”) classes must end in a so-called schwa syllable -*e*, -*el*, -*er*, or -*en* (i.e. phonetically an unstressed [ə] followed by a sonorant from [l r n]).

Strong nouns add -*e* for plural in class NS (e.g. *Arm* ‘arm’, pl *Arme*) and class NU (e.g. *Arzt* ‘physician’, pl *Ärzte*) if the stem itself does not already end in a schwa syllable (e.g. *Kabel* ‘cable’, pl *Kabel*). Class NU furthermore umlauts the stem (i.e. replaces *a*, *o*, *u*, *au* with *ä*, *ö*, *ü*, *äu*, respectively), as does class NR (e.g. *Mann* ‘man’, pl *Männer*), which adds -*er*.

The remaining classes (NM, NWN, and NWS) form their plural in -*n* (e.g. *Schraube* ‘screw’, pl *Schrauben*). The nonnominative singular stem in class NWN (e.g. *Hase* ‘hare’, gen sg *Hasen*) and class NWS (e.g. *Glaube* ‘belief’, gen sg *Glaubens*) is identical with the plural form, while NWN exceptionally adds no -*s* in genitive singular.

All classes except NA build dative plural by adding -*n* to the plural form if it is not already present (e.g. *Männer* ‘men’, dat pl *Männern* but *Hasen* ‘hares’, dat pl *Hasen*).

Figure 1: hierarchy of inflectional schemata



3 Representation of inflectional morphology

Various proposals have been made for the representation of inflectional morphology within constraint-based frameworks like HPSG (cf. Pollard (1994)). We neither adopt a word-syntax approach like that of (Krieger and Nerbonne, 1993) assuming lexical entries for inflectional affixes as well as roots, nor do we make use of lexical rules, as (Meurers and Minnen, 1997) do.

Instead, we follow (Riehemann, 1998) in formulating hierarchically structured schemata of the kind she has developed for derivational morphology but apply them here to inflection and thus carry out a kind of inflectional analysis without lexical rules as projected by (Erjavec, 1996). Our schemata capture inflectional paradigms and can be regarded as relational constraints that relate stems, affixes, and inflected lexical forms.

Figure 1 shows our hierarchy of inflectional schemata, while figure 2 illustrates a concrete schema, namely that for the schwa plural of inflectional class NS. In figure 2 the attribute ftype stands for the inflectional class. The attributes flex, surf, and base represent strings, namely the inflectional ending, surface (i.e. inflected) form, and base form respectively. The symbol @ denotes the reduced vowel [ə] (schwa), and ˜ designates negated values.

Lexical entries are assumed only for basic lexical signs (i.e. uninflected but possibly derived or compounded). Inflected lexical signs result from the interaction of these lexical entries and the inflectional schemata. Figure 3 gives the basic lexical sign (with the omission of feature specifications that are irrelevant for this discussion) for *Hund* 'dog', which is of class NS, followed by the inflected lexical sign for *Hunde* 'dogs', in which the value of the attribute moph (i.e. morphophonology) is an extension of the schema for schwa plural given in figure 2.

The inflectional classes assigned to basic lexical signs are modelled as formal types in the hierarchical structure specified in figure 4. Note that the leaves of this tree correspond exactly to the inflectional classes of German nouns as described above in §2.

Figure 2: schema for schwa-plural (schwa_pl_sort)

$$\begin{bmatrix} \textit{schwa\_pl} & \\ \textsf{ftype} & \textit{ns} \\ \textsf{flex} & @ \\ \textsf{surf} & \textit{atom} \\ \textsf{base} & \textit{atom} \\ \textsf{stem} & \begin{bmatrix} \textit{basic\_ls} & \\ \textsf{moph} & \textit{basic} \end{bmatrix} \\ \textsf{agr} & \begin{bmatrix} \textit{agr} & \\ \textsf{case} & \begin{bmatrix} \textit{case\_avm} & \\ \textsf{ctxt} & \tilde{}\ \textit{dat} \\ \textsf{gen} & \textit{top} \end{bmatrix} \\ \textsf{num} & \textit{plu} \\ \textsf{gend} & \tilde{}\ \textit{fem} \\ \textsf{pers} & \textit{pers} \end{bmatrix} \end{bmatrix}$$

Morphophonemic and morphographemic alternations as shown in nominative plural *Zeit-en* 'times' but *Gabel-n* 'forks' are also covered in our description. Here the realisation of the plural ending *-n* depends on the shape of the noun stem (namely, whether or not it ends in a schwa syllable). In agreement with (Bird and Klein, 1994) and (Erjavec, 1996), we capture such alternations declaratively in a one-level model without recourse to transducers. Our treatment of umlaut adopts part of the techniques of (Trost, 1993).

4 Processing unknown words

In our approach linguistic properties of unknown words are inferred from their sentential context as a byproduct of parsing. After parsing, which requires only a slight modification of standard lexical lookup, lexical entries are appropriately updated. One of our key ideas is a gradual, information-based concept of "unknownness", where lexical entries are not unknown as a whole, but may contain unknown, i.e. potentially revisable, pieces of information (cf. Barg and Walther (1998)). This allows a uniform treatment for the full range of lexical entries from completely known to maximally unknown. As discussed in (Barg and Walther, 1998), our system has been implemented in *MicroCUF*, a derivative of the formalism *CUF* of (Dörre and Dorna, 1993).

Figure 3: feature structures for *Hund* and *Hunde*

$$\begin{bmatrix} \mathit{basic\_ls} & \\ \textsf{moph} & \begin{bmatrix} \mathit{basic} & \\ \textsf{ftype} & \mathit{ns} \\ \textsf{surf} & \boxed{1}\ \mathit{hund} \\ \textsf{base} & \boxed{1} \\ \textsf{agr} & \begin{bmatrix} \mathit{agr} & \\ \textsf{case} & \begin{bmatrix} \mathit{case\_avm} & \\ \textsf{ctxt}\ \tilde{}\ & \mathit{gen} \\ \textsf{gen} & \mathit{case} \end{bmatrix} \\ \textsf{num} & \mathit{sing} \\ \textsf{gend} & \mathit{masc} \\ \textsf{pers} & \mathit{third} \end{bmatrix} \end{bmatrix} \\ \textsf{synsem} & \ldots \end{bmatrix}$$

$$\begin{bmatrix} \mathit{inflected\_ls} & \\ \textsf{moph} & \begin{bmatrix} \mathit{schwa\_pl} & \\ \textsf{ftype} & \boxed{1}\ \mathit{ns} \\ \textsf{flex} & @ \\ \textsf{surf} & \mathit{hunde} \\ \textsf{base} & \boxed{2}\ \mathit{hund} \\ \textsf{agr} & \begin{bmatrix} \mathit{agr} & \\ \textsf{case} & \begin{bmatrix} \mathit{case\_avm} & \\ \textsf{ctxt}\ \tilde{}\ & \mathit{dat} \\ \textsf{gen} & \mathit{case} \end{bmatrix} \\ \textsf{num} & \mathit{plu} \\ \textsf{gend} & \boxed{3}\ \mathit{masc} \\ \textsf{pers} & \boxed{4}\ \mathit{third} \end{bmatrix} \\ \textsf{stem} & \begin{bmatrix} \mathit{basic\_ls} & \\ \textsf{moph} & \begin{bmatrix} \mathit{basic} & \\ \textsf{ftype} & \boxed{1} \\ \textsf{surf} & \boxed{2} \\ \textsf{base} & \boxed{2} \\ \textsf{agr} & \begin{bmatrix} \mathit{agr} & \\ \textsf{case} & \begin{bmatrix} \mathit{case\_avm} & \\ \textsf{ctxt}\ \tilde{}\ & \textsf{gen} \\ \textsf{gen} & \textsf{case} \end{bmatrix} \\ \textsf{num} & \mathit{sing} \\ \textsf{gend} & \boxed{3} \\ \textsf{pers} & \boxed{4} \end{bmatrix} \end{bmatrix} \\ \textsf{synsem} & \boxed{5} \end{bmatrix} \end{bmatrix} \\ \textsf{synsem} & \boxed{5}\ldots \end{bmatrix}$$

Revisable information is further classified as specializable or generalizable, where the former can only become more special, and the latter only more general, with further contexts. Specializable kinds of information include semantic type of nouns, gender, and inflectional class. Among the generalizable kinds of information are the selectional restrictions of verbs and adjectives as well as the case of nouns. Both kinds of information together with nonrevisable (i.e. strict) information can cooccur in a single entry.

The overall approach is compatible with standard constraint-based analyses and makes only a few extra demands on the grammar. Here the revisable information must be explicitly marked as such. Since our model is situated within the framework of typed feature-based formalisms (cf. Carpenter (1992)), revisable information is expressed in terms of formal types. The initial values for revisable information are specified with two distinguished types *u_s* and *u_g* for specializable and generalizable information, respectively. Type unification can be employed for the combination of specializable information, whereas generalizable information requires type union.

The direct combination of revisable information during parsing is unfeasible for various reasons discussed in (Barg and Walther, 1998). It consequently is carried out in a separate step after the current sentence has been parsed. The grammatical analysis itself thus remains completely declarative and only makes use of unification. In order to achieve this separation of analysis and revision we introduce two attributes for generalizable information, namely gen and ctxt, where ctxt receives the information inferred from the sentential context, and gen the potentially revisable information with the initial value *u_g*.

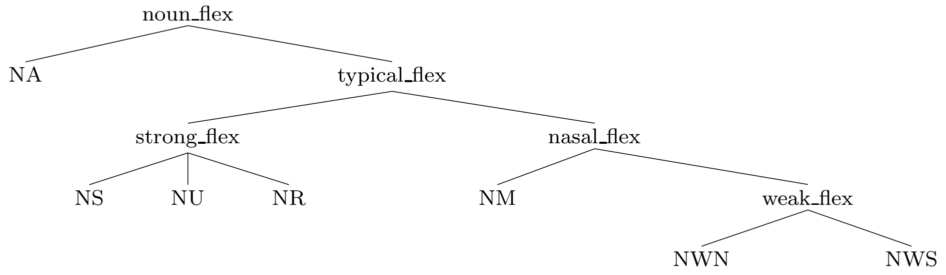
Parsing thus proceeds in an entirely conventional manner, except that lexical look-up for a word with unknown orthography or phonology does not fail but instead yields an underspecified canonical lexical entry. The updating after parsing compares the feature structure of the original lexical entry with that inferred contextually. The specializable information of the former is replaced with the corresponding values of the latter. Moreover, using the attributes gen and ctxt introduced above, the new gen value for generalizable information is computed by the type union of the gen value from the old lexical entry (initialy *u_g*) with the ctxt value resulting from the parse. Actual revision naturally is only carried out when a context in fact provides new information.

5 Incremental inference of inflectional information

In order to process unknown word forms, we postulate canonical lexical entries which are returned by lexical lookup if a word is not recorded in the lexicon. For nouns, this entry corresponds to an underspecified basic lexical sign in which the inflectional class, case, number, and gender are specified with revisable types, i.e. the information can be acquired and updated. Figure 5 shows the basic lexical sign for German nouns (with the omission of feature specifications that are irrelevant for this discussion).

Whereas inflectional class (ftype), number (num),

Figure 4: hierarchy of inflectional types



and gender (gend) are specializable, case is generalizable and hence contains the features gen and ctxt. Note that the initial values for specializable information consist of a disjunction (*;*) of the value *u_s* and the most general appropriate value for the corresponding feature. This ensures the identification of specializable information (via *u_s*) on the one hand, and the correct specializations on the other.

When a sentence containing an unknown noun is parsed, information about the noun comes from different sources: while the surrounding context may supply agreement information, the word form itself together with morphophonological constraints may restrict the possible inflectional class.

As an example we can suppose that the rather infrequent noun *Sund* 'sound', 'strait', which like *Hund* 'dog' belongs to class NS but is unfamiliar to many German speakers, is not recorded in a given lexicon. The class NS contains both masculine and neuter nouns, and these differ in none of their inflected forms. Thus, only agreement information from a context, such as *der enge Sund* 'the narrow strait' (nominative), can establish the gender of *Sund* as being masculine.

Figure 5: feature structure for the underspecified lexical entry

$$\begin{bmatrix} \textit{basic\_ls} & \\ \textsf{moph} & \begin{bmatrix} \textit{basic} & \\ \textsf{ftype} & \textit{(noun\_flex;u\_s)} \\ \textsf{surf} & \boxed{1}\ \textit{atom} \\ \textsf{base} & \boxed{1} \\ \textsf{agr} & \begin{bmatrix} \textit{agr} & \\ \textsf{case} & \begin{bmatrix} \textit{case\_avm} & \\ \textsf{ctxt} & \textit{case} \\ \textsf{gen} & \textit{u\_g} \end{bmatrix} \\ \textsf{num} & \textit{(num;u\_s)} \\ \textsf{gend} & \textit{(gend;u\_s)} \\ \textsf{pers} & \textit{third} \end{bmatrix} \end{bmatrix} \\ \textsf{synsem} & \ldots \end{bmatrix}$$

Even in isolation, the form *Sund* must be singular since its final shape is not compatible with any plural inflection (i.e. it ends neither in *-s* nor in a schwa syllable). Moreover, the morphophonological constraints on stems allow only three possibilities: *Sund* is

- feminine (and then the class is NA, NU, or NM and the case is underspecified)
- nonfeminine and weak (i.e. class NWN or NWS) (and then the case must be nominative)
- nonfeminine and nonweak (and then the case is not genitive)

These hypotheses are captured in the three feature structures depicted in figure 6.

As we have seen, when a word is parsed in context, this provides additional information. If we know, for example, that *Sund* is masculine, the first hypothesis is excluded, and the gender specification of the remaining two hypotheses can be specialized to *masc*. If we additionally encounter *Sund* in dative singular, which is impossible for weak nouns (which must have a final *-n*), then only the third hypothesis remains. Finally, if the plural form *Sunde* occurs the system can specialize the inflectional class exactly to the type NS. The other morphological information cannot be further generalized or specialized, and we have the final lexical entry for *Sund*.

Things are not always this easy. In particular, there may be a number of alternatives both for the segmentation of a form into a stem and an inflectional ending and for the assignment of a stem to a lexeme. Moreover, these alternatives may depend on each other. Thus, the form *Leinen* may be assigned to any of the lexemes *Lein* 'flax' (masc, NS), *Leine* 'rope' (fem, NM), or *Leinen* 'linen' (neut, NS); even in a context, e.g. *Fritz verkauft Leinen* 'Fritz sells ropes/linen', it may be impossible to disambiguate the form. While the nouns *Band* 'book volume' (masc, NU), *Band* 'strip' (neut, NR), *Band* 'bond' (neut, NS, archaic and rare in singular), *Band* 'music band' (fem, NA), and *Bande* 'gang' (fem, NM) may be unlikely to occur all in the same context, they illustrate the dimension of the problems of segmentation and lexical assignment, which in turn con-

Figure 6: hypotheses for *Sund*

$$\left[\begin{array}{ll} \textit{basic\_ls} & \\ \mathsf{moph} & \left[\begin{array}{ll} \textit{basic} & \\ \mathsf{ftype} & \textit{(ns;nu,nm;u\_s)} \\ \mathsf{surf} & \boxed{1}\ \textit{sund} \\ \mathsf{base} & \boxed{1} \\ \mathsf{agr} & \left[\begin{array}{ll} \textit{agr} & \\ \mathsf{case} & \left[\begin{array}{l} \textit{case\_avm} \\ \mathsf{ctxt}\ \textit{case} \\ \mathsf{gen}\ \textit{u\_g} \end{array}\right] \\ \mathsf{num} & \textit{(sing;u\_s)} \\ \mathsf{gend} & \textit{(fem;u\_s)} \\ \mathsf{pers} & \textit{third} \end{array}\right] \end{array}\right] \\ \mathsf{synsem}\ \ldots & \end{array}\right]$$

$$\left[\begin{array}{ll} \textit{basic\_ls} & \\ \mathsf{moph} & \left[\begin{array}{ll} \textit{basic} & \\ \mathsf{ftype} & \textit{(nw;u\_s)} \\ \mathsf{surf} & \boxed{1}\ \textit{sund} \\ \mathsf{base} & \boxed{1} \\ \mathsf{agr} & \left[\begin{array}{ll} \textit{agr} & \\ \mathsf{case} & \left[\begin{array}{l} \textit{case\_avm} \\ \mathsf{ctxt}\ \textit{case} \\ \mathsf{gen}\ \textit{(nom;u\_g)} \end{array}\right] \\ \mathsf{num} & \textit{(sing;u\_s)} \\ \mathsf{gend} & \textit{(}\tilde{}\ \textit{fem;u\_s)} \\ \mathsf{pers} & \textit{third} \end{array}\right] \end{array}\right] \\ \mathsf{synsem}\ \ldots & \end{array}\right]$$

$$\left[\begin{array}{ll} \textit{basic\_ls} & \\ \mathsf{moph} & \left[\begin{array}{ll} \textit{basic} & \\ \mathsf{ftype} & \textit{(}\tilde{}\ \textit{nw;u\_s)} \\ \mathsf{surf} & \boxed{1}\ \textit{sund} \\ \mathsf{base} & \boxed{1} \\ \mathsf{agr} & \left[\begin{array}{ll} \textit{agr} & \\ \mathsf{case} & \left[\begin{array}{l} \textit{case\_avm} \\ \mathsf{ctxt}\ \textit{case} \\ \mathsf{gen}\ \textit{(}\tilde{}\ \textit{gen;u\_g)} \end{array}\right] \\ \mathsf{num} & \textit{(sing;u\_s)} \\ \mathsf{gend} & \textit{(}\tilde{}\ \textit{fem;u\_s)} \\ \mathsf{pers} & \textit{third} \end{array}\right] \end{array}\right] \\ \mathsf{synsem}\ \ldots & \end{array}\right]$$

stitute part of the more general problem of disambiguation in natural language processing. We have no magic solution for the latter, but in our approach such examples must be handled with disjunctive representations until the context provides the necessary disambiguating information.

6 An alternative model using finite-state techniques

Alternatively, the incremental identification of inflectional types can be modelled within the framework of finite-state automata (cf. Sproat (1992)) without recourse to unification-based grammar formalisms. A FSA can be defined that has an alphabet consisting of vectors specifying the stem shape and ending (and thus the segmentation) as well as the agreement information of possible word forms. Starting in an initial state corresponding to the constraints that apply to all unknown words, the FSA is moved by successive forms of an unknown lexeme together with their agreement information into successor states that capture the incrementally accrued inflectional information. The FSA may reach a final state, in which case the inflectional class has been uniquely identified, or it may remain in a nonfinal state. A lexicon would simply record the latest state reached for each noun.

Implementation of this model is greatly complicated by the problems of disambiguation just discussed in §5. In general, the states of the FSA must capture disjunctions not only of inflectional classes, but also of segmentation and gender alternatives. The application of automatic induction techniques to corpora appears to be essential, and we are currently pursuing possibilities for this.

7 Conclusion

We have taken the inflection of German nouns to illustrate a general type-based approach to handling unknown words and the incremental accrual of their lexical information. The techniques can be applied not only to other classes of inflected words and to other languages, but also to other aspects of lexical information such as the valency of verbs. This may allow practical systems for natural language processing to be enhanced so as to utilize input information that otherwise is discarded as noise.

References

Petra Barg and Markus Walther. 1998. Processing unknown words in HPSG. In *Proceedings of COLING-ACL'98*, pages 91–95, Montreal.

Steven Bird and Ewan Klein. 1994. Phonological analysis in typed feature systems. *Computational Linguistics*, 20:455–491.

Michael R. Brent. 1991. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of 29th ACL*, pages 209–214, Berkeley.

Lynne Cahill and Gerald Gazdar. 1999. German noun inflection. *Journal of Linguistics*, 35:1–42.

Robert Carpenter. 1992. *The Logic of Typed Feature Structures.* Cambridge University Press.

Harald Clahsen. 1999. Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences*.

Jochen Dörre and Michael Dorna. 1993. CUF – a formalism for linguistic knowledge representation. In Jochen Dörre, editor, *Computational Aspects of Constraint-Based Linguistic Description*. IMS, Universität Stuttgart. Deliverable R1.2.A, DYANA-2 – ESPRIT Project 6852.

Karen Ehrlich and William J. Rapaport. 1997. A computational theory of vocabulary extension. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pages 205–210.

Tomaž Erjavec. 1996. *Unification, Inheritance and Paradigms in the Morphology of Natural Languages*. Unpublished doctoral dissertation, University of Ljubljana.

Udo Hahn, Manfred Klenner, and Klemens Schnattinger. 1996. Learning from texts - a terminological meta-reasoning perspective. In S. Wermter, E. Riloff, and G. Scheler, editors, *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pages 453–468. Springer, Berlin.

James Kilbury, Petra Barg, and Ingrid Renz. 1994. Simulation lexikalischen Erwerbs. In Sascha Felix, Christopher Habel, and Gerd Rickheit, editors, *Kognitive Linguistik: Repräsentation und Prozesse*, pages 251–271. Westdeutscher Verlag, Opladen.

Jean-Pierre Koenig. 1999. *Lexical Relations*. CSLI Publications.

Hans-Ulrich Krieger and John Nerbonne. 1993. Feature-based inheritance networks for computational lexicons. In Ted Briscoe et al., editor, *Inheritance, Defaults, and the Lexicon*, pages 90–136. Cambridge University Press, Cambridge.

Detmar Meurers and Guido Minnen. 1997. A computational treatment of lexical rules in HPSG as covariation in lexical entries. *Computational Linguistics*, 23:543–568.

Martin Neef. 1998. A case study in declarative morphology: German case inflection. In Wolfgang Kehrein and Richard Wiese, editors, *Phonology and Morphology of the Germanic Languages*, pages 219–240. Max Niemeyer Verlag, Tübingen.

Carl Pollard and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago University Press.

Susanne Riehemann. 1998. Type-based derivational morphology. *Journal of Comparative Germanic Linguistics*, 2:49–77.

Richard Sproat. 1992. *Morphology and Computation*. MIT Press, Cambridge, Mass.

Harald Trost. 1993. Coping with derivation in a morphological component. In *Proceedings of 6th EACL*, pages 368–376.

Markus Walther and Petra Barg. 1998. Towards incremental lexical acquisition in HPSG. In *Proceedings Joint Conference on Formal Grammar, Head-Driven Phrase Structure Grammar, and Categorial Grammar*, Saarbrücken.

Dieter Wunderlich. 1999. German noun plural reconsidered. Manuscript, University of Düsseldorf.