Detecting Mistakes or Finding Misconceptions? Diagnosing Morpho-Syntactic Errors in Language Learning

Wolfgang Menzel
Universität Hamburg, Department Informatik
Vogt-Kölln-Straße 30, D-22527 Hamburg, Germany
email: menzel@informatik.uni-hamburg.de

Abstract

Diagnosing morphosyntactic errors in a language learner's utterance can substantially profit from a direct treatment of morphological regularities if wordform-internal relationships are integrated into the general constraint-based model of correctness. The paper describes model structures for a range of selected problems from different inflected languages, namely German, Russian and Bulgarian, and discusses the additional diagnostic capabilities, which are made available by a description, which is based on morphemes instead of full-forms.

1 Introduction

Over the last two decades a number of systems have been devised, which try to provide language learners with individualized feedback to help them in acquiring the regularities of a new language. By means of an interactive learning environment more or less constrained utterances are elicited from the learner and techniques from Computational Linguistics are applied to analyse them and diagnose possible errors.

Even if the target language is fairly rich in morphological regularities, morphology is hardly ever addressed by such systems (c.f. (Schwind 95; Heift 98; Reuer 03)). If it is considered at all, this mostly happens in the context of lexical knowledge acquisition, where the system offers its resources for inspection by the student rather than attempting to diagnose learners utterances (c.f. (Nerbonne *et al.* 98; tenHacken & Tschichold 01)). It turns out, however, that integrating the regularities of the inflectional system into the syntactic model for a properly defined subset of a language and trying to find out, which of the combined regularities have been violated by the language learner, might reveal deeper insights into which misconceptions led to an error. Eventually, more useful feedback can be produced, helping the student to overcome the problem. To achieve this goal, word forms need to be decomposed into their morphological building blocks and constraints on how to combine these elementary parts into legal word forms and well-formed utterances have to be supplied. By making these constraints defeasible, error hypotheses can be simulated in order to obtain an error explanation that seems most plausible for the given diagnostic problem. Such an approach has been successfully applied to the morpho-syntactic regularities which hold between the word forms of a (partial) natural language utterance (Menzel 88b; Menzel 88a; Menzel 92). This paper describes extensions which are necessary to also include word-internal relationships.

2 Model-based diagnosis

The diagnostic approach used in this paper is model-based (deKleer *et al.* 92; Menzel 06). It comes with the decisive advantage that only knowledge about the conditions under which an utterance can be considered correct is represented. This avoids the necessity to anticipate a vast amount of conceivable erroneous constructions, a task which must be considered almost open-ended even for rather simple types of exercises.

Modelling syntactic regularities by means of constraints is rather straightforward for phenomena like the agreement between different word forms in a natural language utterance. Given an ill-formed utterance error hypotheses can be derived by simulating constraint violations and propagating their consequences through the constraint net. A set of constraint violations is considered a plausible diagnosis of the utterance if it

- explains the mistake in the sense that the properties of the student's input are entailed by the assumed constraint violations together with the remaining correctness conditions,
- and is minimal in the sense that no proper subset of constraint violations also explains the error.

The decision on a plausible diagnosis often cannot be taken by only considering locally available information. There are usually a number of alternative and competing diagnosis candidates, which represent different views on the erroneous situation. Perhaps the simplest case of such a diagnostic ambiguity is depicted in Figure 1, where three word forms of the English sentence *These fish is small.* have to agree with respect to their number feature. Unfortunately, they do so pairwise, but not if taken together. The necessary condition for this phenomenon to occur is the ambiguity of the noun *fish* which can be both, singular or plural.

This ambiguity is reflected in two competing correction proposals, either into singular *This fish is small.*, or into plural *These fish are small.*, which result from the simulation of constraint violations in a slightly modified model (cf. Figure 2). It reflects another view on the student, assuming that she observed the underlying rules of the language (namely the two agreement conditions) but instead ignored the morpho-syntactic features of the respective word forms as defined in the dictionary. Therefore, the model imposes strict agreement between all three forms but allows the diagnosis to assume a mismatch between the true feature values of a word form and what the student believes them to be. Accordingly, two diagnoses can be derived under which a consistent interpretation can be established: either the student might have assumed *these* to be singular or *is* to be plural. Deciding between these two possibilities eventually will require to take knowledge about the situational context of the utterance into consideration.

Languages with a more sophisticated inflectional system than the English one confront a learner with more challenging networks of well-formedness constraints. Figure 3 shows the correctness conditions of a German prepositional phrase consisting exactly of a preposition, a determiner, an adjective and a noun.

While such a model allows the student to freely compose her solution from among the available lexical entries, it still imposes relatively hard restrictions, by requiring all four word forms as being mandatory ones. This limitation can be somewhat weakened by including explicit constraints on linear precedence ('lp') and optionality into the model: word forms are considered optional as long as not stated otherwise. Thus, the extended model in Figure 4 requires

- at least one of adjective or noun to be present ('noun ∨ adjective'),
- to have a preposition either as such or as part of a phonologically contracted form together with the determiner ('preposition `exor` preposition+determiner'), and
- the existence of an adjective whenever a graduating particle has been used ('graduating particle → adjective').

Moreover, it makes the presence of the preposition and the determiner mutually dependent on each other ('preposition ↔ determiner').

Despite its apparent simplicity such a model provides all the necessary means to diagnose a wide range of erroneous student utterances in a controlled exercise environment. It also raises the question whether an application to languages with an even richer inflectional system might be promising. Suprisingly, the constraint net for noun phrase inflection in a language like Russian looks even simpler than its German counterpart (cf. the left part of Figure 5 for a model with just a single adjective).

The apparent reason for this simplicity is the lack of determiners and the highly regular inflectional system, where all adjectives in a noun phrase carry the very same type of ending. A comparison with Bulgarian shows the same tendency, lacking even a fully developed system of case markings (cf. the right part of Figure 5). Instead, a definiteness marker can be attached to either the noun or the adjective and a corresponding constraint has to ensure that it can only be used with the first component of a noun phrase ('only-det-initial'). Since by default the adjective is optional in the student solution, the constraint prohibits the noun to carry a definiteness marker as long as there is a preceding adjective (cf. Table 1).

Unfortunately, in some cases a simple constraint set does not allow us to derive maximally informative diagnoses. In a noun phrase like

*най-сърдечен благопожелания
(the most heartful wishes)

the information supplied by the diagnosis is rather shallow, simply stating the mismatch in gender and number, with най-сърдечен being masculine singular, but благопожелания neuter plural. Another possible explanation, namely the

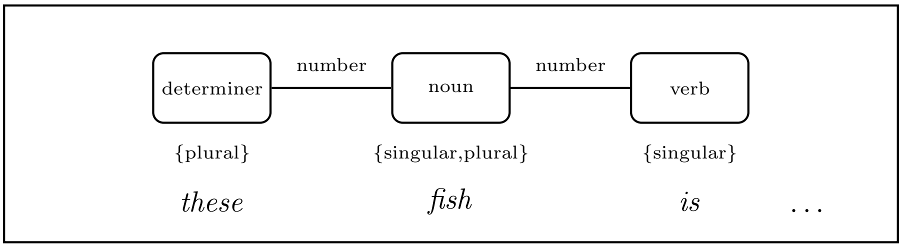

Figure 1: A rule-based model for a simple agreement problem in English

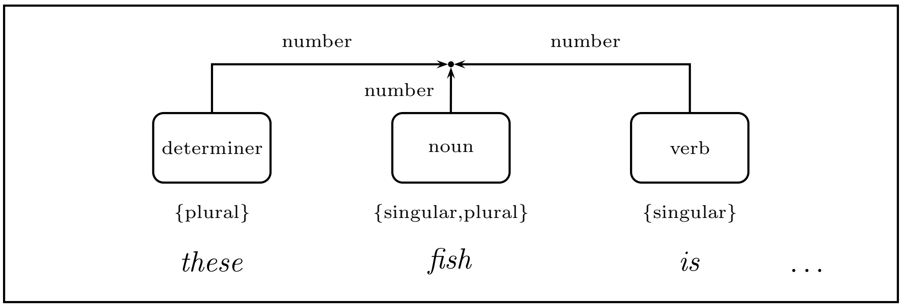

Figure 2: A lexically-based model for the simple agreement problem from Figure 1

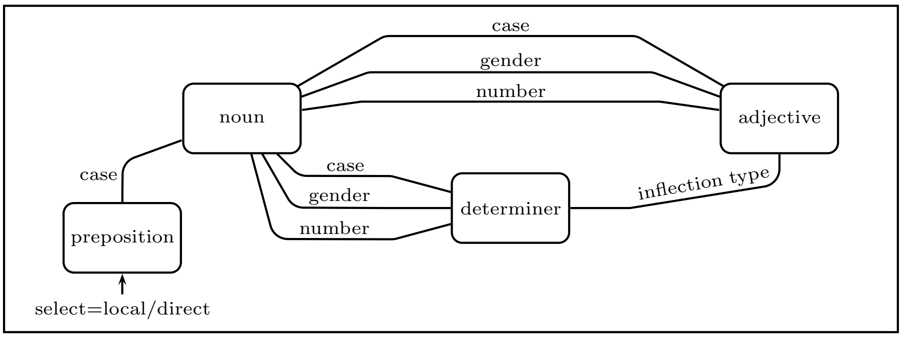

Figure 3: Agreement and government conditions for a German prepositional phrase

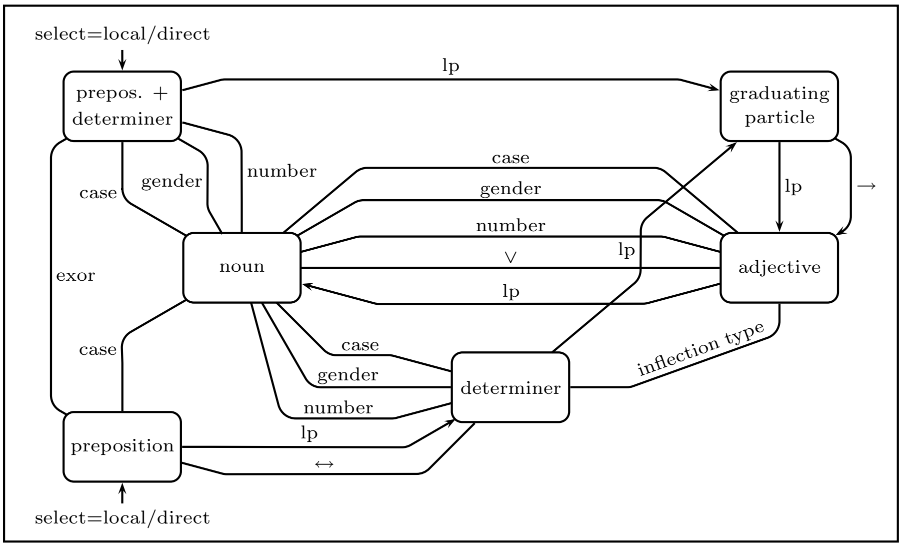

Figure 4: Well-formedness conditions for a German prepositional phrase

student might have considered благопожелания a masculine singular noun because of the superficial similarity with the ending of a word form like съдия (male judge) is not available to the current modelling.

3 Constraint-based models of inflectional morphology

A constraint-based diagnosis works best, if the problem description, which can be derived from the student solution, is rich enough in information. This relationship seems to be universal and has already been observed in completely different domains, cf. (Kodaganallur *et al.* 05) for an application of constraint-based modelling to statistical hypothesis testing or (Le & Menzel 06), who diagnose solutions for simple logical puzzles. As a consequence, exercises should be designed, as to elicit as much information as possible from the student. There is, however, no easy way to get access to more information in a language learning setting without either rendering the exercises unnatural from a communicative point of view (i.e. by asking for grammatical features directly), or making them too complex, thus introducing a degree of ambiguity, which not only incurs an efficiency problem but also spoils the diagnostic results with a huge number of interpretation possibilities.

Fortunately, there is an alternative way to provide the diagnostic component with more detailed information about the student's solution by breaking down word forms into morphemes and supplying the system with a description of their morpho-syntactic properties. Attempting to analyse correct input utterances one certainly would choose a sequential arrangement of two components with an isolated morphological component preceding the syntactic parser. If, however, a diagnosis of ungrammatical input is aimed at, both kinds of linguistic regularities need to be integrated into a single model. This integration can be achieved by describing the possible combinations within and between word forms by means of constraints. For that purpose roots and inflectional endings have to be categorized into distributional classes (paradigms) and a kind of agreement between them has to be enforced, so that e.g. only noun endings can be attached to noun stems and only those endings are considered, which belong to the inflectional paradigm required by the root. In the same manner other open class word forms can be dealt with. Moreover, the approach can be extended to consider additional requirements, like a wordform-internal gender or number agreement.

Figure 6 shows an example for the inflection of a very simple German noun phrase consisting of a determiner and a noun. Since determiners are rather opaque from a morphological point of view, only the noun is split into root and inflectional ending and the compatibility between them is established by a set of gender-specific paradigms. Additionally, there needs to be a constraint on number agreement since some German nouns exhibit root inflection (Umlaut: Apfel$_{sg}$ → Äpfel$_{pl}$). Both, the root and the inflectional ending of the noun have been made obligatory, i.e. zero endings need to be modelled explicitly. This option has been chosen since even a zero ending might contribute morpho-syntactic features which are essential for the diagnosis. Alternatively, the ending could be considered optional and if not provided by the student, the features of the zero ending are assumed as default values.

Here and in all the subsequent models the linear precedence relation between a root and the corresponding inflectional ending has been modelled directly. Since the likelihood that it might ever be violated in a student's solution is fully negligible, it could, of course also be captured implicitly by the variable binding procedure of the diagnosis.

The major contribution of this enhanced modelling is the additional possibility to also diagnose non-words, which have been constructed according to assumptions about false analogies in the language to be learned. Such an error occurs for instance, if the student is mislead by a superficial similarity when trying to infer the inflection of a yet unfamiliar word from another one probably acquired earlier. Take for example the two German words *Apfel* (apple) and *Schachtel* (box). Although they share the same word final rhyme, they have different genders (masculine and feminine respectively) and therefore follow different inflectional paradigms. Misled by the similarity a student might be tempted to derive the plural form **die Apfeln* in analogy to the plural form *die Schachteln*. This, however, is not a valid word form of German. Nevertheless, the diagnosis is able to precisely pinpoint the underlying reason for the error by hypothesizing an *internal* gender

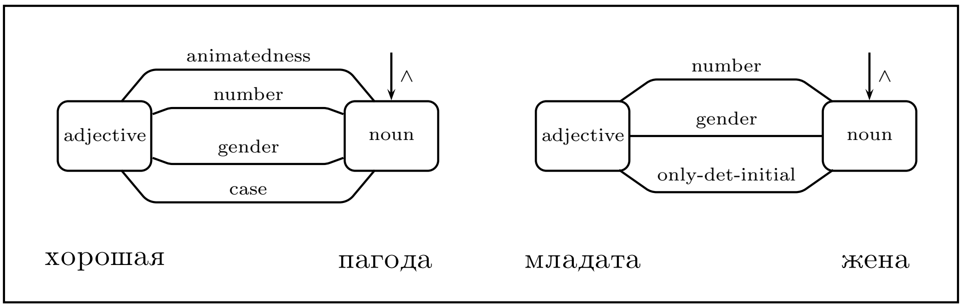

Figure 5: Well-formedness constraints for noun phrases in Russian (left) and Bulgarian (right)

	only at adjective	determiner at adjective and noun	only at noun	no determiner
with adjective	+	−	−	+
without adjective			+	+

Table 1: Truth-values of the constraint `only-det-initial`

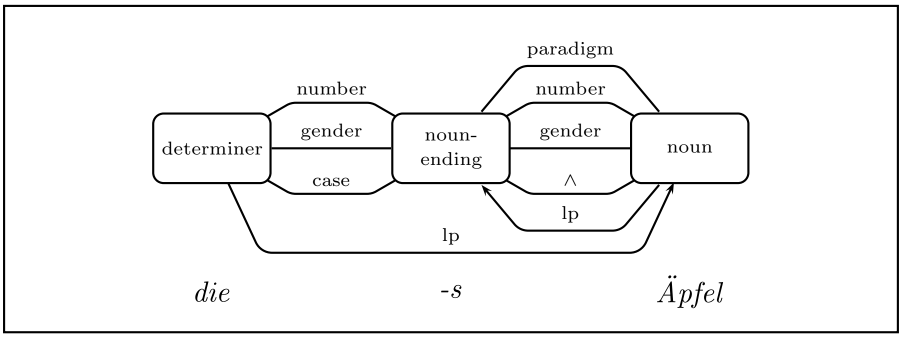

Figure 6: Morpheme-based model for a German noun phrase

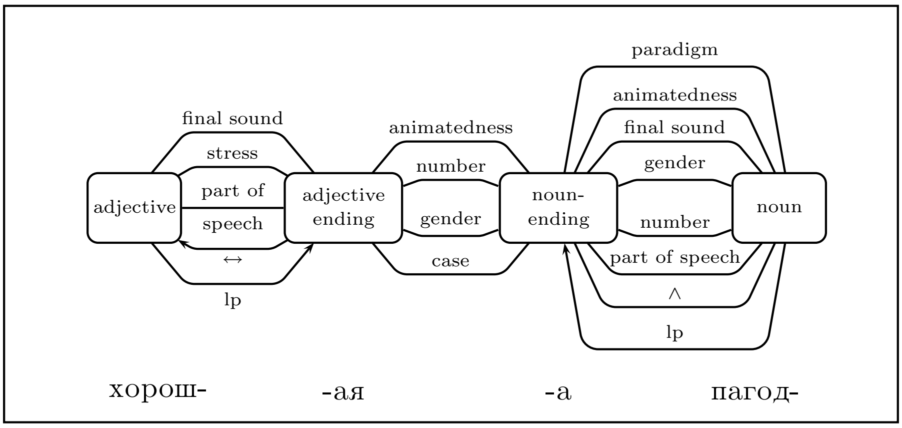

Figure 7: Morpheme-based model of Russian noun phrase inflection

disagreement between the stem and the ending of the noun as being the most simple and therefore rather plausible error explanation: _’Apfel’_ is masculine, not feminine.

Ignoring stem inflection is a another possibility for producing non-words from German nouns. This is an error type, which can be observed in learners utterances quite frequently. Starting from the existing knowledge about the dative plural form of _die Gabel_ being _(mit) den Gabeln_, the student might conclude that _der Apfel_ is inflected according to the very same pattern, namely _*(mit) den Apfeln_. Although _Apfeln_ is also a non-word in German, the diagnosis still succeeds in determining that the inflectional paradigm for _Apfel_ requires umlaut and it is exactly this property that the student’s solution is lacking.

Applying the principle of morphological decomposition to the models of Russian and Bulgarian noun group inflection as given in Figure 5 introduces a host of new constraints and gives rise to the expectation that this additional model information might eventually translate into a corresponding improvement of the diagnostic quality.

The model for Russian (Figure 7) specifies two additional types of word form internal constraints which capture phonological regularities of the Russian declension, namely

- the difference between root stress and final stress for adjectives (’stress’), i.e. новый vs. большóй , (new/big) and
- the general dependence of the vowel’s ending on the preceding consonant being a hard or a weak one (’final sound’), i.e. маленький vs. новый (small/new).

The adjective and its ending are optional but their existence is mutually dependent on each other (↔), whereas the noun and its ending are obligatory. Zero endings are treated the same way as in the German model.

Due to the explicit modelling of these regularities the diagnosis is now in a position to also detect phonologically inappropriate inflectional forms of Russian, e.g. *большый дом (big house), and to compute the quite plausible hypothesis about a word form internal agreement error for the stress constraint. It can be explained to the student by a feedback message like ”большой has word stress on the ending, which therefore must change from -ый to -ой”. Similarly, for the phonologically ill-formed *маленькый дом (small house) an internal agreement violation for the final sound constraint is detected and verbalized e.g. as ”The root маленьк- ends with a hard consonant but the ending -ый can only be used with a soft one.”

With the treatment of the (optional) definiteness marker the model for Bulgarian introduces yet another specific feature. This marker, which in Bulgarian is attached to the end of the adjective or the noun, is modelled by means of two variables (’adjective definite’ and ’noun definite’). An additional constraint (→∼) ensures that whenever there is a definiteness marker at the adjective it must not appear on the noun.

In the above mentioned utterance *най-сърдечен благопожелания the inflectional ending -я is ambiguous between masculine singular and neuter plural. Therefore, the agreement constraints between the endings can be satisfied locally, guiding the system to the alternative diagnosis of an internal agreement error between the root and the ending of the noun with respect to gender: ”-я is not an appropriate singular marker for a neuter noun like благопожелание”. Along this line of reasoning an error explanation has been derived, which appears to be much closer to the underlying misconception, compared to a full form based diagnosis, which simply states a conflict on a much shallower level.

4 Conclusions

Decomposing a full form into its morphemes and specifying their compatibility conditions as defeasible constraints brought us into a position to precisely diagnose agreement errors which are caused by the superficial similarity of morphological items. Such a similarity always results in an ambiguity and alternative error hypotheses become available to the diagnosis component. It then can check, which one of the possible interpretations leads to a simpler, hence more plausible diagnosis. Compared to the agreement violations found for a full form model, the corresponding diagnostic results much closer resemble the underlying reasons which might have caused the error.

In addition, this ability to explicitly reason about the ambiguity introduced by individual morphemes also makes possible the analysis of

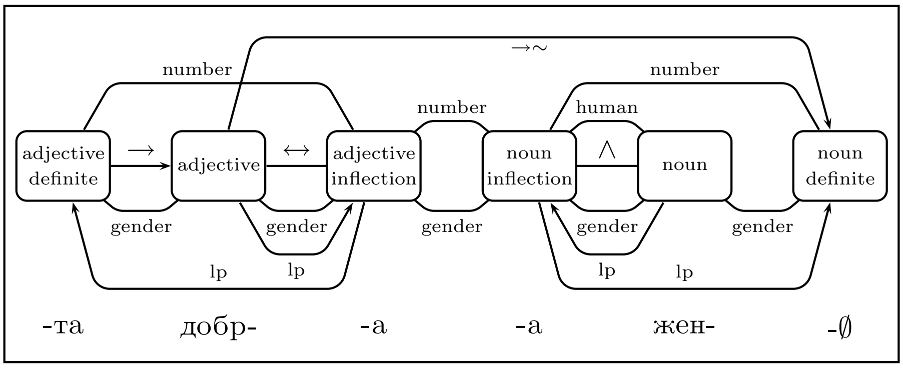

Figure 8: (Simplified) morpheme-based model of Bulgarian noun phrase inflection

non-words, which have been composed out of valid morphemes but are assembled in a way that violates some of the wordform-internal compatibility constraints.

Although they do without any kind of error anticipation, constraint-based models have been shown to provide a surprisingly high degree of diagnostic precision for limited exercise types like the formation of a German prepositional phrase (Menzel 92). This precision even allows to derive one or several correction proposals for a given student solution by simply replacing the erroneous items by appropriate alternatives, which can be retrieved from the dictionary, directly using the feature values determined by the diagnosis as the ones, which establish global consistency with minimal effort. Therefore, the system is even able to provide the student with a repair suggestion, transforming e.g. the repeatedly used erroneous Bulgarian example *най-сърдечен благопожелания into its corrected version

Най-сърдечни благопожелания!

Acknowledgements

I am grateful to Marina Marinova for explaining me the fundamentals of Bulgarian morphology .

References

(deKleer *et al.* 92) J. de Kleer, A. K. Mackworth, and R. Reiter. Characterizing diagnoses and systems. *Artificial Intelligence*, 56, 1992.

(Heift 98) Trude Heift. *Designed Intelligence: A Language Teacher Model.* Unpublished PhD thesis, Simon Fraser University, 1998.

(Kodaganallur *et al.* 05) V. Kodaganallur, R. R. Weitz, and D. Rosenthal. A comparison of model-tracing and constraint-based intelligent tutoring paradigms. *International Journal on Artificial Intelligence in Education*, 15:117–144, 2005.

(Le & Menzel 06) Nguyen Thinh Le and Wolfgang Menzel. Constraint-based problem generation for a self-assessment system. In *Proc. 14th Int. Conf. on Computers in Education, Workshop on Problem Authoring, -Generation, and -Posing in a Computer-Based Learning Environment*, Bejing, 2006.

(Menzel 88a) Wolfgang Menzel. Diagnosing grammatical faults - a deep-modelled approach. In *Artificial Intelligence III, Proceedings of AIMSA '88*, pages 319–326, Amsterdam, 1988. North Holland.

(Menzel 88b) Wolfgang Menzel. Error diagnosing and selection in a training system for second language learning. In *Proc. 12th Int. Conf. on Computational Linguistics*, pages 414–419, Budapest, 1988.

(Menzel 92) Wolfgang Menzel. *Modellbasierte Fehlerdiagnose in Sprachlehrsystemen*, volume 24 of *Sprache und Information.* Niemeyer, Tübingen, 1992.

(Menzel 06) Wolfgang Menzel. Constraint-based modelling and ambiguity. *International Journal on Artificial Intelligence in Education*, 2006.

(Nerbonne *et al.* 98) John Nerbonne, Duco Dokter, and Petra Smit. Morphological processing and computer-assisted language learning. *Computer-Assisted Language Learning*, 11(5):421–437, 1998.

(Reuer 03) Veit Reuer. *PromisD: Ein Analyseverfahren zur antizipationsfreien Erkennung und Erklärung von grammatischen Fehlern in Sprachlehrsystemen.* Dissertation, Humboldt-Universität zu Berlin, 2003.

(Schwind 95) Camilla B. Schwind. Error analysis and explanation in knowledge based language tutoring. *Computer Assisted Language Learning*, 8(4):295–324, 1995.

(tenHacken & Tschichold 01) Pius ten Hacken and Cornelia Tschichold. Word manager and call: Structured access to the lexicon as a tool for enriching learners' vocabulary. *ReCALL*, 13:121–131, 2001.