GIVING PROSODY A MEANING

Christian Lieske[4] *Johan Bos*[1] *Martin Emele*[2] *Björn Gambäck*[3] *CJ Rupp*[1]

(1) Computational Linguistics, University of Saarland; Postfach 151150; D-660 41 Saarbrücken
Tel: +49 681 302 4679, Fax: +49 681 302 4351, {bos,cj}@coli.uni-sb.de
(2) Institute of Computational Linguistics, University of Stuttgart; Azenbergstrasse 12; D-70174 Stuttgart
Tel: +49 711 121 1372, Fax: +49 711 121 1366, emele@ims.uni-stuttgart.de
(3) Centre for Speech Technology, Royal Institute of Technology; S-100 40 Stockholm
Tel: +46 8 790 8884, Fax: +46 8 790 7854, gamback@speech.kth.se
(4) Computer Science Department, Swiss Federal Institute of Technology; CH-1015 Lausanne
Tel: +41 21 693 2589, Fax: +41 21 693 5278, lieske@di.epfl.ch

ABSTRACT

Systems for spoken-language understanding can use prosodic information on the speech recognition side as well as the linguistic processing side. In the former case, prosody improves recognition accuracy and speed. In the latter case, it contributes to the computation of meaning. Interfacing prosodic processing to language analysis has so far been mainly concerned with speeding up the parsing process. The actual integration of prosodic information into the semantic part of a language understanding system, or into the transfer part of a translation system, has mostly been left aside.

We describe how prosody has been used in the syntactic-semantic and transfer modules of the Verbmobil spoken dialogue translation system. On the syntactic-semantic side, prosody is currently used for the solution of three different problems: insertion of clause boundaries, selection of sentence mood (declarative, question, etc), and assignment of semantic focus. On the transfer side, the prosodic information is allowed to influence the lexical choice of the system.

1. INTRODUCTION

Systems for spoken-language understanding can use prosodic information on the speech recognition side as well as the linguistic processing side. In the former case, prosody improves recognition accuracy and speed. In the latter case, it contributes to the computation of meaning. The following paragraphs discuss this meaning-related use of prosody in the spoken-language machine translation system Verbmobil (VM).

The overall goal of the VM system is to provide speech translations from both German and Japanese to English [8, 10]. The scenario assumes that two businessmen, one native speaker of German, the other a native speaker of Japanese, try to schedule an appointment. Both the dialogue partners possess at least a passive knowledge of English, so that the conversation may proceed mostly in English. In case one of the users' active English knowledge turns out to be insufficient, the user may switch to his/her native language using the Verbmobil system. The system then translates the user's mother tongue speech into spoken English.

The "Verbmobil-Forschungsprototyp 1.0" (released in October 1996) comprises 43 different modules which use a specifically designed architecture and protocol to communicate with each other [5]. The end-to-end speech translation rate in the domain is 74.2%, with a vocabulary size of about 2500 words in the German-to-English subsystem and about 400 words in the Japanese-to-English part. The system works at an average speed of 5.7 times real time.

The prosody module of the Verbmobil system connects to the recording unit and the speech recognizer, on the input side, and the morphology module, on the output side. The data structure for communication (input from the recognizer and output) are *word lattices* whose edges are annotated with recognition probabilities and so-called *infostrings* which amongst others may encode three different kinds of prosodic information: sentence modality, phrase boundaries, and stress [7].

In a very wide sense, Verbmobil comprises five modules related to linguistic processing: syntactic-semantic processing (SynSem), semantic evaluation, transfer (TR), generation, and dialogue. Of those, currently only SynSem and TR make use of prosodic information, so we restrict our sketch of the system to these two components.

Syntactic-semantic processing is based on a parser for a unification-based grammar which interleaves syntactic analysis and semantic construction [3]. Semantic construction compositionally builds representations called VITs (Verbmobil Interface Terms) which include semantic, syntactic, pragmatic and prosodic information. VITs allow the representation of ambiguities such as the relative scope of quantifiers, and are very flat (minimal recursive) structures [4].

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

Verbmobil adheres to the idea of transfer-based machine translation (transfer based on semantic representations to be specific), i.e., there is a non-trivial mapping between the structures resulting from the analysis of a source-language (SL) utterance and the structures used for the generation of the corresponding target-language (TL) expression. The approach is compositional, meaning that the semantic predicates for the SL are mapped in chunks onto the semantic predicates for the TL [6]. Transfer rules can utilize all of the information found in the VIT (e.g., values of syntactic features and prosodic cues). Two features are worth mentioning: Since pragmatic information is found in the VITs, the rules are even able to do some local anaphora resolution. Secondly, since VITs are underspecified semantic representations, transfer can preserve ambiguities, e.g., related to scopal relationships, when mapping from SL to TL.

The usage of prosody in the SynSem module will be further discussed in Section 3, while Section 4 addresses the same issue within the transfer module. First, however, we will describe how the VM language modules are interfaced to the prosodic processing.

2. PROSODY AND LANGUAGE

The integration of prosodic information into the syntactic and semantic analysis poses both practical and conceptual problems. Thus, we had to integrate prosody values into what is essentially, still, a modified string parser. In addition, the probabilistic values provided by the prosody analysis had to be adapted to the feature domain of unification-based linguistics, dealing essentially in discrete symbols.

The first problem was addressed by a simple expedient of introducing an additional symbol into the parser string after each lexical word. These *prosodic word forms* (proswofs) provide a simple channel for the prosodic information associated with each edge in the word lattice. Similarly, each pre-terminal rule in the grammar has an additional category in its expansion. This gives a rule of the form `c -> c p`, where `c` is a lexical category and `p` the category `proswof`. An example of such a prosodic word form might have the form `ak10b3gr3prsfrage`, following a word with no stress accent at the end of a phrase with question intonation.

Adjusting the prosodic information for the normal feature domain of the grammar incurs a certain loss of information since the values provided represent weights on the acceptability of a certain analysis, actually taking the form of inverse logarithmic probabilities. Either the probability of each of the possible analyses is provided or both negative and positive probabilities for a particular feature. In the logically oriented domain or unification-based linguistics only discrete judgements are possible, either no information is available or a definitive analysis is made. Hence, even the conversion of 2-place decimal probability to a numeric feature value implies that the input value must be in some way *normalised.* For example, the information provided for sentence mood in the word lattice gives weights for each of three possibilities: statement, progradient or question, e.g., `(M 0.99 0.34 0.23)`. The proswof given above results from taking the feature value corresponding to the minimum of these values, `frage` (question).

3. PROSODY AND MEANING

The syntactic-semantic processing module makes *conservative use* of prosodic information in the sense that it is allowed to influence the analysis only if syntactic and semantic evidence do not override it. If, e.g., the grammar does not allow for a clause boundary but prosody indicates that there is one, then the grammar takes precedence. Currently, prosodic information is dealt with in three areas of SynSem: segmentation, sentence mood and focus.

3.1 Segmentation

The input to a spoken dialogue system is structured in terms of turn-taking in the dialogue. Syntactic and semantic analyses can be assigned to meaningful linguistic entities at the clausal or phrasal level, but first the turn has to be segmented into a sequence of linguistically credible segments. The prosodic indications of clausal boundaries are crucial to this process, but they are also probabilistically generated, so it may be necessary for the syntactic constraints to override the available prosodic information. The interaction between prosodic and syntactic constraints in the TUG parser [3] for German has been reported in more detail in [2, 9].

The syntactic analysis effectively defines the distribution of both optional and obligatory phrase boundaries. In the latter case a syntactically licensed boundary marker would be introduced if the prosodic marker was absent. The distribution of such boundary markers was determined by a number of factors, including parser efficiency and corpus coverage, as well as purely syntactic constraints. The effect of such segmentation on translation result can be demonstrated by the distribution of the initial particle *ja* in German, as in the following examples.

ja bei mir geht prinzipiell jeder Montag **SynPB**
Well, as far as I'm concerned, in principle every
Monday is possible.

ja **SynPB** *das paßt mir* **SynPB**
Yes. That suits me.

where **SynPB** marks a phrase boundary.

3.2 Sentence Mood

In many cases syntactic or semantic criteria are sufficient to determine whether a sentence is declarative, imperative or interrogative. An obvious example would be the occurrence of a topicalised wh-phrase as syntactic evidence of a question. We can also apply simple semantic considerations to determine that certain mental states are not appropriate content for an imperative in normal discourse, to the extent that modal verbs often lack morphological imperative forms. However, clause structure in German does not always determine sentence mood. The normal verb-second word order found in main clauses may be, formally, ambiguous between statement and question and verb initial clause according to standard grammar may be imperative or interrogative. In addition, spoken language typically throws up alternative constructions or isolated phrases which may be ambiguous with respect to mood. In particular, verb initial statements can arise through topic drop, rendering the sentence mood, in principle, three ways ambiguous.

Prosodic information often provides a clear guide to the resolution of such ambiguities. Typical examples from the Verbmobil corpus include:

Machen	*Sie*	*einen*	*Vorschlag*
make	you	a	proposal

Will you make a proposal? with rising intonation or *Make a proposal!* with progradient intonation.

Kommen	*Sie*	*zu*	*mir*	*ins*	*Büro*
come	you	to	me	in the	office

Will you come to my office? or *Come to my office!*

3.3 Focus

Prosodic information is also used to help determine the focus value of focus sensitive adverbs, such as *auch* or *nur*, determined by the occurrence of a stress accent on a relevant word within the scope of the operator. Compare the following translation examples, where words with stress accent are in bold face.

Das paßt auch bei **mir**
That suits me **too**.

das *paßt* **auch** *bei mir*
That **also** suits me.

In the current Verbmobil system it is not yet possible for the Generation module to communicate the stress patterns to synthesis, although this would of course be required for adequate pronounciation of the translations of utterances involving focus adverbs.

4. PROSODY AND TRANSFER

Transfer makes use of prosodic information when processing focus related adverbs and adjectives (e.g., *noch*). For this to work, the symbolic representation of prosodic information as described in the previous section needs to be interfaced with the transfer rules. The general form of a transfer rule is given by

```
SLSem,SLConds TauOp TLSem,TLConds.
```

where `SLSem` and `TLSem` are sets of semantic entities in the source and target languages, respectively. `TauOp` is an operator indicating the intended application direction (one of `<->`,`->`,`<-`), while `SLConds` and `TLConds` are optional sets of SL and TL conditions. For a more detailed description of the transfer formalism the reader is referred to [6].

The main difference between `SLSem` and `SLConds` is that the former is matched against the input and replaced by the `TLSem`, whereas the conditions act as filters on the applicability of individual transfer rules without modifying the input representation. Hence `SLConds` may be viewed as general inferences which yield either true or false depending on the context. In the examples to follow the relevant context is a test whether a certain predicate is accentuated or not. This test is implemented via an abstract interface predicate `pros_accent` which accesses the prosodic information found in the VIT.

Rule (1) and (2) provide an example of *lexical choice* for the focussing adverb *noch.*

```
(1) [L:noch_fadv(F,S), L1:indef(I,G,S1)],
      [pros_accent(L)] ->
                [L:another(I,G,S1)].
```

```
(2) [L:noch_fadv(F,S)],
      [L:indef(I,G,_)] -> [].
```

Rule (1) says that the predicate related to `noch` and the predicate `indef` related to the indefinite article `a` should be mapped to a predicate for `another` if the predicate `noch` is stressed. The relevant test is checked in the conditional part of the transfer rule via the above mentioned interface predicate `pros_accent`. It tests whether the predicate labelled

L, e.g., noch_fadv carries a prosodic accent mark in the prosody slot of the VIT. In case of absence of stress for noch, this rule will not trigger and the default rule (2) is applied, simply deleting the noch predicate.

For example an utterance like *Können wir noch einen Termin ausmachen?* is translated either as *Could we arrange* **another** *appointment?* or as *Could we arrange* **an** *appointment?* depending on whether *noch* is stressed or not.

It is interesting to note that many of the discourse and focus particles do not have a direct counterpart and hence may be deleted in the English translation if they are not stressed (cf. also [1]).

Another example where prosodic information might be useful is the ambiguity of German ein between the indefinite article and the cardinal one, e.g., *einen Tag* could either mean *a day* as in *Wir brauchen einen Tag im May* (we need a day in May) or *one day* as in *Es dauert einen Tag* (it lasts for one day) depending on the context. Information about stress on the indefinite pronoun might be used as an heuristic for choosing the cardinal reading if no further discourse knowledge is available to make a definite decision.

In the semantic representation for German the ambiguity is underspecified by using the vague predicate ein_card_qua which allows for both possibilities. Since there exists no equivalent predicate for English, transfer has to decide which reading is more plausible. The rule in (3) maps the underspecified predicate to the cardinal reading whereas in case of absence of stress the default rule (4) maps it to the indefinite article.

```
(3) [L:ein_card_qua(I,R,S)],
      [pros_accent(L)] ->
                  [L:udef(I,R,S),L1:card(I,1)].
(4) [L:ein_card_qua(I,R,S)] ->
                  [L:indef(I,R,S)].
```

5. FURTHER WORK

Our research agenda for the coupling of prosodic processing and linguistic processing addresses two issues: Firstly, we want to identify additional cases where prosody can add constraints pertaining to syntax or semantics. Secondly, we are aiming at refining the calculations that are involved in the addition of constraints due to prosodic information.

ACKNOWLEDGEMENTS

This research was funded by the VERBMOBIL project under grant number 01 IV 101 R from the German Federal Ministry for Education and Research (BMBF).

We would also like to acknowledge the contribution of various colleagues at a number of institutions, most notably Anton Batliner, Ralf Kompe and Elmar Nöth at the University of Erlangen–Nuremberg; Hans–Ulrich Block, Manfred Gehrke, Stefanie Schachtl and Tobias Ruland of Siemens Corporate Research in Munich; Karsten Worm and Yoshiki Mori at the University of Saarbrücken; and Michael Dorna at the University of Stuttgart.

REFERENCES

[1] J. Alexandersson and B. Ripplinger, Disambiguation and Translation of German Particles in Verbmobil, VM Memo 70, DFKI GmbH Saarbrücken, Saarbrücken, Germany, 1996.

[2] G. Bakenecker, U. Block, A. Batlinger, R. Kompe, E. Nöth, and P. Regel-Brietzmann, *Improving Parsing by Incorporating 'Prosodic Clause Boundaries' into a Grammar*, Proc. 3rd ICSLP, pp. 1115–1118, Yokohama, Japan, 1994.

[3] H. U. Block and S. Schachtl, *Trace and Unification Grammar*, Proc. 14th Int. Conf. on Computational Linguistics, pp. 658–664, Nantes, France, 1992. ACL.

[4] J. Bos, B. Gambäck, C. Lieske, Y. Mori, M. Pinkal, and K. Worm, *Compositional Semantics in Verbmobil*, Proc. 16th Int. Conf. on Computational Linguistics, pp. 131–136, København, Denmark, 1996. ACL.

[5] T. Bub, W. Wahlster, and A. Waibel, *Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation*, Proc. 22nd ICASSP, München, Germany, 1997. IEEE.

[6] M. Dorna and M. C. Emele, *Semantic-based Transfer*, Proc. 16th Int. Conf. on Computational Linguistics, pp. 316–321, København, Denmark, 1996. ACL.

[7] W. Hess, A. Batliner, A. Kiessling, R. Kompe, E. Nöth, A. Petzold, M. Reyelt, and V. Strom, Prosodic Modules for Speech Recognition and Understanding in Verbmobil, In Y. Sagisaka *et al* eds., *Computing Prosody: Approaches to a Computational Analysis and Modelling of Prosody of Spontaneous Speech*, pp. 363–384. Springer, New York, 1996.

[8] M. Kay, J. M. Gawron, and P. Norvig, *Verbmobil: A Translation System for Face-to-Face Dialog*, Lecture Notes#33. CSLI, Stanford, California, 1994.

[9] R. Kompe, A. Kießling, H. Niemann, E. Nöth, A. Batliner, S. Schachtl, T.Ruland, and U. Block, *Improving Parsing of Spontaneous Speech with the Help of Prosodic Boundaries*, Proc. 22nd ICASSP, München, Germany, 1997. IEEE.

[10] W. Wahlster, *VERBMOBIL: Translation of Face-to-Face Dialogs*, Proc. 3rd EUROSPEECH, pp. 29–38, Berlin, Germany, 1993.