From Information Structure to Intonation: A Phonological Interface for Concept-to-Speech

Hannes Pirker, Georg Niklfeld, Johannes Matiasek and **Harald Trost**[+]
{hannes,georgn,john,harald}@ai.univie.ac.at
Austrian Research Institute for Artificial Intelligence (OFAI)*
Schotteng. 3, A-1010 Vienna, Austria
[+]Department of Medical Cybernetics and Artificial Intelligence University of Vienna
Freyung 6, A-1010 Vienna, Austria

Abstract

The paper describes an interface between generator and synthesizer of the German language concept-to-speech system VieCtoS. It discusses phenomena in German intonation that depend on the interaction between grammatical dependencies (projection of information structure into syntax) and prosodic context (performance-related modifications to intonation patterns).

Phonological processing in our system comprises segmental as well as suprasegmental dimensions such as syllabification, modification of word stress positions, and a symbolic encoding of intonation. Phonological phenomena often touch upon more than one of these dimensions, so that mutual accessibility of the data structures on each dimension had to be ensured.

We present a linear representation of the multidimensional phonological data based on a straightforward linearization convention, which suffices to bring this conceptually multilinear data set under the scope of the well-known processing techniques for two-level morphology.

1 Introduction

The task of interfacing between a tactical generator and a speech synthesizer is two-fold: A grammatical description enriched with semantic and pragmatic features has to be translated into a (qualitative) phonological description which then has to be mapped onto the set of (quantitative) parameter values needed as input to the synthesizer.

The requirements imposed by a concept-to-speech system differ from those on both text generation and text-to-speech systems. In text generation the generator produces a sequence of abstract descriptions of word forms which are–either by direct access to a lexicon or via a morphological component–transformed into strings of graphemes and output. With concept-to-speech the task is more complex. Not only is segmental information influenced by morphonology and post-lexical rules (covering, e.g., reduction and assimilation phenomena) but–more important–suprasegmental information must be provided as well.

Compared to text-to-speech the task is at the same time easier and more difficult. Information from pragmatic, semantic and syntactic layers are readily available. This eliminates the need to analyze an input text for necessary cues to come up with proper pronunciation and prosody. On the other hand all this information must be properly accounted for to come up with an adequate description of the utterance that–when fed into the synthesizer–produces high-quality output. In particular, pragmatic-semantic features must be mapped onto (abstract) prosodic features.

We employ an extended version of two-level morphology (Trost 91) for this interface.[1] The formalism proved to be very well suited for the task. The various almost independent subsystems can be kept conceptually separate resulting in good transparency while at the same time enabling the necessary amount of interaction between them.

2 A Concept-to-Speech Generation System

Our concept-to-speech generation system consists of a pipeline of modules (Fig. 1). A text

* This work has been sponsored by the *Fonds zur Förderung der wissenschaftlichen Forschung (FWF)*, Grant No. P10822.

[1] The extension regards the fact that the system allows the use of (feature-based) external information–so-called filters–to restrict the application of two-level rules.

planning component produces sentence plans, which are fed into the tactical generator.

The implementation basis for the tactical generator is the FUF (Elhadad 91) system. FUF is based on the theory of functional unification grammar and employs both phrase structure rules and unification of feature descriptions. Input is a partially specified feature description which constrains the utterance to be generated. Output is a fully specified feature description (in the sense of the particular grammar) subsumed by the input structure, which is then linearized to yield a sentence.

The tactical generator has two layers. One is dealing with sentence level generation, producing a tree-like description of a sentence, the leaves of which are lemmata annotated with morphosyntactic and prosodic features. The second performs generation at the word level producing annotated phonological representations of the inflected word forms which are fed into the extended [2] two-level phonology component applying morphological and phonological rules to arrive at the representation used as input for speech synthesis.

A distinguishing feature of the grammar used in the generator is the integration of sentence-level and word-level processing within the same formalism.

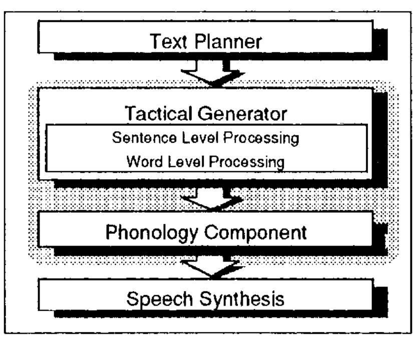

Figure 1: Architecture

This architecture forms an ideal platform for the implementation of the phonological interface. Necessary adaptions are limited to the data used: An existing grammar was extended with features describing the information structure. The lexicon consists of entries in phonemic form (using SAMPA notation) enriched with information like (potential) accent and syllable boundary positions.

Input to the synthesizer is a SAMPA string enriched with qualitative encodings of prosodic information (e.g., pitch accent, pauses, ...) produced by the two-level rules. Phonological specifications of intonation are processed by a phonetic interpreter (Pirker et al. 97) that transforms these qualitative labels into quantitative acoustic parameters. Although some interpretative work is done within the synthesizer, no linguistically motivated transformations are supposed to take place there. These all are performed within the two-level component.

3 The Phonological Interface

3.1 Phenomena handled

The phonological description in extended two-level morphology – in our case rather two-level *phonology* –serves as the central interface where the modules for grammar processing and for speech synthesis meet and communicate.

A fairly complex model of phonology is required in the system, also because the overall objective of the project was to investigate whether and how conditions in the concept-to-speech task favour a more elaborate treatment of prosodic parameters in speech generation.

The phonological description is implemented in the extended two-level framework described in section 2 and works over a lexicon of phonemic (rather than graphemic) representations of word stems and inflectional affixes. Morphotactic processing is thus restricted to inflection, whereas compounding and derivational affixation are encoded in the lexicon, which is typically small in domain-tailored concept-to-speech systems.

Nevertheless, in segmental phonology, the component must compute morphonological rules in inflection as well as post-lexical rules which interact with syllabification and cliticization.

To determine German syllabification and cliticization correctly, it is necessary to operate on structures larger than single words. Therefore phonological processing applies to chunks whose size depends on the one rule in the system that requires the largest phonological context to operate correctly. Because of the intonation rules discussed in section 4, phonological

[2]The filter handling uses the FUF formalism and the same unification machinery as the grammar.

processing applies to the whole utterance.

The three phonological aspects segmental representation, syllabification, and word stress are mutually dependent in German phonology in all logically possible directions (Niklfeld et al. 95). The phonology component treats them in a unified description, which also covers the rare cases of word-internal and phrase-level stress shift in German.[3]

While some segmental and supra-segmental rules in the phonological description depend on phonological context only, some others (like the rule for stress shifts as described above) depend on grammatical information on levels as high up as textual representation. For example, the German word for "weather" loses word stress in compounds when they appear in weather-reports (where the concept weather is "textually exophoric" (Benware 87)). Such phenomena are encoded in our extended two-level system by phonological rules which access the grammatical representation via feature-filters.

There are few theoretical frameworks in computational linguistics for tackling such a breadth of phonological issues. Linguistically ambitious approaches are often designed with little regard to ease of use in large descriptions, whereas leaner formalisms do not scale well to complex data stretching across a number of phonological dimensions. The chosen framework of extended two-level phonology stands between these poles.

3.2 Linearization of multi-tier phonological structures

As the two-level framework assumes one lexical and one surface string only, we use a linear representation of our multidimensional phonological data, as follows:

Each linear phonological string in the component stands for a multi-tier structure which combines a given number of separate dimensions of phonological structure. The tier of phonological segments (members of the German SAMPA set) is used to provide the backbone of skeletal points on which all units of the representation are linked together. Each unit on any phonological tier has scope over/has as its domain a continuous section of skeleton points. For each tier, a convention is provided which designates that part of each domain that is used for the linking. For some supra-segmental tiers (syllables, phonological words) the leftmost unit of the scope domain as computed by the respective rule is used for this purpose. For other tiers the domain edges are unspecified in the lexicon (stresses and accents, which have scope over stretches of syllables), and therefore other well-defined parts of the scope domain are used for the linking (such as the vocalic nucleus of a syllable). Where it appears natural to do so, units on certain phonological tiers are also linked to right domain edges (as is the case with phrase and boundary tone markers, which have scope over any phonological material between a nuclear tone and the right boundary of an intonation phrase.)

While these representations clearly encode some fragment of autosegmental phonology in an implicit way, they do not allow for the attachment of more than one suprasegmental unit from the same tier to a single segmental unit. Such power was not needed in our application.

The representation allowed for easy incremental extensions to our descriptions, as additional tiers of representation were added as the coverage of higher-level prosodic issues such as sentence intonation was extended.

3.3 Implementational notes

Using the linearized representation, the well-known processing schemes for two-level morphology can be applied directly. Contemporary compilers for two-level morphology allow to specify sets of symbols that are ignored in individual rules. Extensive application of such syntactic sugar enables us to keep the rule formulations over the collapsed representation economical and relatively transparent. We note in passing that although collapsing multilinear data-structures onto a single tier increases the likeliness of combinatorial explosion in processing when using the two-level automata as transducers, it turns out that in our already quite complex description this does not become a real problem.

In earlier publications, we described how we implement phonological generalizations that stretch across phonological dimensions (Niklfeld et al. 95), and we proposed implementations of suprasegmental issues such as stress shift and

[3] Otherwise, German has lexically specified word stress.

the projection of pitch accents depending on focus information (Niklfeld & Alter 96). We have also discussed time structure (Alter et al. 96). In section 4 we go beyond this to show that intonation in German has properties that are best implemented by combining our two-level phonological description, which is well-suited to express constraints on linear contexts, with the power of a unification-based feature grammar.

4 Dealing with Intonation

This section describes the novel approach of using the extended two-level component for specifying "appropriate" intonation and phrasing.

4.1 Different perspectives

The diversity of factors that influences intonation is mirrored in the variety of research that deals with intonation:

Phonologists and phoneticians are concerned with the inspection of the form of intonation contours, while on the other hand there is a strong tradition in the field of syntax (keyword: focus projection) and semantics/pragmatics (keyword: given vs. new information) that merely deal with the problem of accent *location*, neglecting its form.

Another strand of research deals with the coupling of information structure and phonology, i.e., the tight association of meanings and tunes such as in (Prevost & Steedman 94) where the classification of the utterance's elements along the dimensions *theme/rheme* and *focus/ground* unambiguously triggers the selection of tones.

In the field of text-to-speech synthesis, at last, intonation most often is handled by using algorithms and heuristics that intermingle information on syntax, punctuation, word-class information etc. in a rather unstructured way.

4.2 Our design

In our system a strict separation of levels is employed: only the two-level component deals with tonal specifications. Within the tactical generator only *candidate positions* for both pitch accents and phrasal boundaries are selected.

This reflects the fact that though prosody heavily depends on grammatical and pragmatic factors, its realization is also strongly influenced by phonological and phonetic constraints which are much more "naturally" handled by the two-level component. In the terminology of two-level morphology the grammar provides a underspecified *lexical* representation from which the concrete *surface form* is derived. In the lexicon every (accentable) word contains an abstract pitch tone (T) within its phonemic representation. The "lexical boundaries" (B), i.e., candidates for boundaries between intonational phrases (IP), are inserted by the generator in between words and these T and B are then mapped to GToBI labels (*German Tones and Break Indices* — (Grice et al. 96)) or discarded i.e., mapped it to surface 0.

The following example (in pseudo-code) defines a basic condition on the IP: it contains at least one, at most three pitch accents, and has an obligatory boundary tone.

```
<IP>        ::= {<PitchTone>{<PitchTone>}}
                 <PitchTone><IP_Bound>
<IP_Bound> ::= L-L% | L-H% | H-L% | H-H%
<PitchTone>::= <RisingT> | <FallingT>
<RisingT>   ::= H* | L+H* | L*+H
<FallingT> ::= L* | H+L* | H+!H*
```

In order to determine the realization of a T the grammatical information the generator provided for the word in question is inspected via the filter mechanism: E.g. if a words was marked as unaccented (acc -) the tone will be discarded or the selection of boundary tones is triggered by the sentence type (L-L% in the case of assertions):

```
T:0    <=  _ filter:(head (phon (acc -)));
B:L-L% <=> _ filter:(head (s-type assert));
```

While the rules discussed so far have been pure filter applications the last rule encodes a constraint on phonological context:

```
B:L-H% => <FallingT> <UnaccSyll>* _ |
    <RisingT> <UnaccSyll> <UnaccSyll>+ _;
```

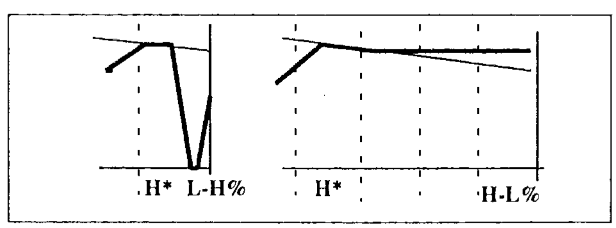

Figure 2: Contours to be **avoided** (vertical lines designate syllable boundaries)

The rationale behind this rule is, that we want to avoid the contours shown in figure 2 when realizing IP boundaries. The L-H% boundary basically designates a fall-rise contour which should be a felicitous if the last pitch accent before the boundary was a falling one. The second term states, that after a rising pitch accent the same boundary contour is to be produced only if the pitch peak is followed by two or more unaccented syllables thus ensuring that there is "enough time" to produce the fall-rise. At the same time the production of the concurring H-L% is blocked, which would produce a long monotonous stretch on a high level, that might be perceived as unnatural.

The rules thus also implement some of the variability in prosody that is due to the interaction of phrasing and pitch accents much in the spirit of tone-linking (Gussenhoven 84).

5 Conclusion

With our approach we unify some of the efforts outlined in 4.1 and come up with a system that is more clearly structured than the "algorithmic" approach.

By basing our work on GToBI -- and thus on a variant of Pierrehumbert's model on intonation -- we have access to the wealth of phonological research undertaken in the tone sequence paradigm.

The handling of accentuation and phrasing by the generator resembles the syntacto-semantic approaches. Only a few tags such as emphasis [EMPH] and (conceptual or textual) givenness [GIVEN] which are rather easily identifiable by the conceptual component and have a straightforward influence on the phonetic realization are used. In this respect our approach is less refined than, e.g., (Prevost & Steedman 94) as no fully fledged semantic module is integrated that could deal with aspects of information structure in a really principled way

On the other hand we employ a very flexible and transparent phonological model. But not all intonation contours that can be observed in human speakers are equally convenient for the use in synthetic speech, where the deviations in duration, amplitude, etc. may lead to results that are perceived as highly unnatural. We thus restrict the set of possible contours licensed by the GToBI to a simplified subset.

The system is implemented and deals with the task of generating monologuous weather reports.

References

Alter K., Matiasek J., Niklfeld G.: Modeling Prosody in a German Concept-to-Speech System, in Gibbon D.(ed.), Natural Language Processing and Speech Technology, Mouton de Gruyter, Berlin, 1996.

Benware W.A.: Accent Variation in German Nominal Compounds of the Type (A (BC)), Linguistische Berichte, 108:102-27, 1987.

Elhadad M.: FUF: The Universal Unifier User Manual, Dept.of Computer Science, Columbia University, 1991.

Grice M., Reyelt M., Benzmüller R., Mayer J., Batliner A.: Consistency in Transcription and Labelling of German Intonation with GToBI, Proc. of ICSLP 96, Philadelphia, pp.1716-19, 1996.

Gussenhoven C.: On the grammar and semantics of sentence accents, Dordrecht: Foris, 1984.

Niklfeld G., Pirker H., Trost H.: Using Two-Level Morphology as a Generator- Synthesizer Interface in Concept-to-Speech, in Proc. of Eurospeech 95, Madrid, 2:1223-26, 1995.

Niklfeld G., Alter K.: Covering prosody in concept-to-speech via an extended two-level-phonology component, in Computational Phonology in Speech Technology – 2nd Meeting of SIGPHON, Santa Cruz, CA, 1996.

Matiasek J., Trost H.: An HPSG-Based Generator for German - An Experiment in the Reusability of Linguistic Resources, in Proc. of COLING 96, Copenhagen, pp.752-57, 1996.

Pirker H., Alter K., Matiasek J., Trost H., Kubin G.: A System of Stylized Intonation Contours for German, in Proc. of Eurospeech 97, Rhodes, Greece, 1:307-10, 1997.

Prevost S., Steedman M.: Specifying Intonation from Context for Speech Synthesis, Speech Communication, 15:139-153, 1994.

Trost, H.: X2MORF: A Morphological Component Based on Augmented Two-Level Morphology, in: IJCAI-91, Morgan Kaufmann, San Mateo, CA, pp.1024-1030, 1991.