gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbmobil
Verbundvorhaben

# Diagnostic Evaluation in Linguistic Word Recognition


Julie Carson-Berndsen and
Martina Pampel

Universität Bielefeld



Vm

**Report 38**
August 1994

August 1994

Julie Carson-Berndsen and
Martina Pampel

Universität Bielefeld (UBI)
Fakultät für Linguistik und Literaturwissenschaft
Universitätsstr. 25
Postfach 10 01 31
33501 Bielefeld

Tel.: (0521) 106 - 3519/11
e-mail: {berndsen, martina}@asl.uni-bielefeld.de

**Gehört zum Antragsabschnitt:** 15.6 Interaktive Phonologische Interpretation

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter dem Förderkennzeichen 01 IV 101 B 2 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

# Contents

**Abstract**

This report is concerned with a new method of evaluation for the Linguistic Word Recognition component of the Verbmobil-Project: *Architektur*. A two stage model of diagnostic evaluation is presented consisting of logical and empirical evaluation steps. Logical evaluation is carried out according to a data model which acts as optimal input in order that each component participating in the evaluation process can be tested for soundness and completeness. Inconsistencies can thus be remedied before empirical evaluation of the model is undertaken using real data. The diagnostic evaluation method has been operationalised within the Bielefeld Extended Evaluation Toolkit for Lattices of Events (BEETLE).

# 1 Introduction

This report is concerned with a new approach to evaluation which has been developed in connection with Linguistic Word Recognition in the *Verbmobil Project 15.6 Interactive Phonological Interpretation*. In the sections below, it is demonstrated how current evaluation procedures are not sufficient for catering for the area of Linguistic Word Recognition. Instead a new diagnostic approach to evaluation of individual components within a spoken language recognition system is presented which takes soundness and completeness issues into account. The standard evaluation procedure allows for evaluation at the word and sentence level. However, it is claimed here that evaluation is necessary at all levels of recognition and that although current *Verbmobil* evaluation software [13] (termed Braunschweig evaluation in the rest of this report) may be adapted to cater for evaluation at the syllable and phoneme level, the constraint that the output lattices of the component must contain at least one connected path imposes a restriction upon the components of Linguistic Word Recognition which has negative effects and which stands in opposition to the aims of *Interactive Phonological Interpretation*.

In *Verbmobil* Project 15.6 *Interactive Phonological Interpretation*, word recognition is performed by applying linguistic knowledge below the word level. Based on recent developments in the areas of phonology and morphology, a more flexible approach to word recognition is followed which is based on the notion of events [1, 3, 5, 10, 11]. The motivation for the application of the event concept in the area of linguistic word recognition concerns the projection problem at the phonetics/phonology interface [6]. In Project 15.6 temporal relations between phonological events form the basis of a grammar which allows a projection of a finite set of actual structures (i.e. in the corpus) onto an infinite set of potential structures allowing for the treatment of new syllables and words. A nonconcatenative, compositional approach to phonology and morphology based on autosegmental tiers of events avoids a rigid segmentation at the phonetics/phonology interface and thus coarticulation effects (overlap of properties) can be described. A decision on segmentation into phonological or morphological units can be postponed by underspecifying the autosegments both temporally and in terms of their features until sufficient information is available. The Linguistic Word Recognition component supplies linguistic knowledge constraints which have been used for optimising stochastic systems [12].

The aim of Linguistic Word Recognition is, by integrating stochastic and linguistic-symbolic approaches with fine granularity, to achieve corpus-independent and speaker-independent speech recognition. BELLEx3 is part of an experimental development environment in which components can demonstrate differing interaction strategies and parameter settings for different modes of analysis. The free parameterisation of the system allows for linguistically adequate parameters to be chosen which define a compromise between maximal word recognition rates and minimal analysis overhead.

The Linguistic Word Recognition Component in Project 15.6 consists of two components; a syllable parser (SILPA) and a morphoprosodic parser (MORPROPA) which together with an acoustic event recogniser (HEAP, Universit"at Hamburg) form the components of the BELLEx3 word recognition system. Each component must be evaluated individually in order to assess the value of the system as a whole. This report is concerned with the evaluation of the syllable parser and the morphoprosodic parser. These two components produce three output lattices which are relevant for evaluation: a phoneme lattice, a syllable lattice and a word lattice. The phoneme lattice can be regarded as a side-effect of syllable recognition as it is possible to derive the phoneme lattice top-down from the phonemes which occur in the recognised syllables. It is important to note, however, that it is not syllable or phoneme hypotheses which are passed from SILPA to MORPROPA but rather underspecified subsyllable events and that syllable and phoneme lattices are generated soley for the purposes of syllable evaluation.

The output of the Linguistic Word Recogintion component as a whole is a word hypothesis lattice which differs from a connected word graph to the extent that the connectedness condition is not a necessary requirement. The output lattice allows the existence of overlapping hypotheses and gaps between hypotheses. Although overlap relations and gaps between hypotheses can be interpreted as precedence relations (i.e. with the help of an absolute overlap parameter or value relative to which the degree of overlap is defined), the formal criterion for connectedness in the sense of [14] is not fulfilled.

## 2 Problems of Evaluation in Linguistic Word Recognition

As mentioned in the previous section, the output of Linguistic Word Recognition is a word hypothesis lattice which differs substantially from the connected word graph in that the connectedness condition is not a necessary condition. Clearly it is possible to construct a connected word graph on the basis of the word lattice output by artifically splitting phonological events into phonemic units using temporal statistics and mapping the lattice to a chart in the sense of [7]. However, the method preferred in Linguistic Word Recognition corresponds to a delayed level-specific segmentation which allows overlaps and gaps between hypotheses. Example 1 provides an example of an overlap of word hypotheses which must be arbitrarily segmented if a connected word graph is to be constructed.

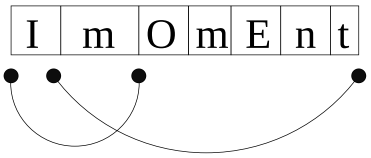


Figure 1: Example of overlapping word hypotheses

The phonogical event 'm' belongs to two word hypotheses although its temporal duration does not correspond to the realisation of two separate phonemic segments 'mm'. An arbitrary splitting into two phonological events 'm' and 'm' based on temporal statistics is therefore unreliable. This case, in particular, shows the argument against an arbitrary segmentation of temporally overlaping word hypotheses to be convincing.

In connection with evaluation as practiced in Verbmobil, the evaluation procedure cannot be directly applied to the output of the components of Linguistic Word Recognition due to the fact that the existence of at least one connected path through the output lattice is not guaranteed. In order to be able to apply the evaluation procedure, the components of Linguistic Word Recognition have two possible strategies for interpreting overlapping hypotheses (cf. figure 2) for the construction of a connected graph; either hypothesis **a** can be interpreted as preceding hypothesis **b** (i.e. option (i) in figure 3) or only hypothesis **c** can be chosen (i.e. option (ii) in figure 3).

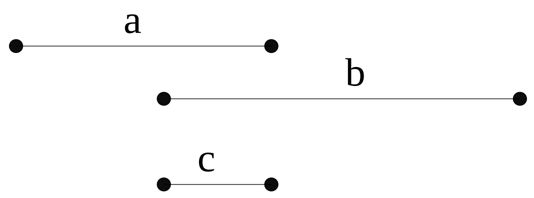


Figure 2: Overlapping Hypotheses

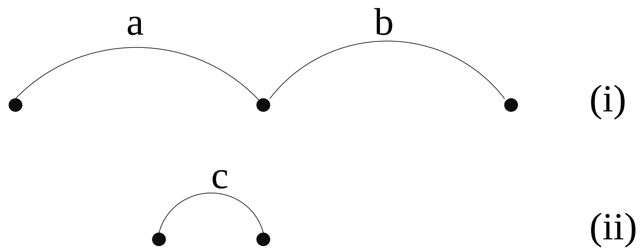


Figure 3: Mapping to Chart

Stochastic word recognition models provide the n-best hypotheses with respect to some particular threshold in the form of a connected word graph. However, due to the fact that the aim of project 15.6 is to examine to what extent linguistic knowledge can be applied in word recognition, this type of output is not desirable, since this condition imposes a non-linguistic restriction on the output format. Gaps may occur in the output for the following reasons:

- hesitations, pauses, silence, incomplete words and sentences.
- underspecification in the input data due to the fact that information is missing in the signal; although it is not possible to complete this information bottom-up, it may be possible to do so top-down.
- incorrect input data from the point of view of the linguistic component; constraint relaxation may be employed here.
- the linguistic knowledge base may be incomplete.

The Linguistic Word Recognition component caters for gaps between hypotheses by allowing underspecified event and feature structures in the output. Inflectional endings, for example, are not easily recognised and therefore congruence features can be left underspecified until a definite decision can be made based on further bottom-up or top-down information.

In this paper we make the claim that an evaluation procedure must take signal endpoints into consideration when defining the notion of connectedness. A certain (maximal) amount of overlap between units must be allowed and they may still be considered to be connected. Reference files must therefore contain signal annotations for the utterances, and not merely the utterance as a string. A disadvantage of our approach is that reference files must be generated for all utterences to be evaluated (i.e. regardless of whether the same utterance is spoken by different speakers or several utterances are spoken by the same speaker).

The Braunschweig evaluation procedure, as employed in Verbmobil has been described in [13]. This evaluation procedure has been developed on the basis of experience with conventional word recognition systems which produce connected word graphs as output and for this task, it is very suitable. Input to the evaluation procedure is a word lattice (which is in fact a connected word graph) and evaluation is carried out with respect to the best path through the lattice. However, in addition to the connectedness condition, the evaluation procedure requires a lexicon which covers the relevant corpus. If the recognised units in the lattice are not contained in the lexicon, a message to this effect is produced. Since Linguistic Word Recognition claims to be able to cater for new words, it clearly does not make sense to use an evaluation procedure which stipulates that the recognised units must be contained in the accompanying lexicon. Linguistic Word Recognition, on the other hand, distinguishes internally between *actual* units which occur in the relevant corpus and *potential* units which are wellformed according to the grammar. For the evaluation of the components of Linguistic Word Recognition, it was also found useful to have a visualisation tool for the output lattices. This is described in more detail below.

The first set of data which was evaluated in Linguistic Word Recognition (all utterances from CD-ROM PhonDat Vol. II (1992) 'Zugauskunft' by the speaker SATD, referred to below as the ASL-Scenario) was that agreed upon by *Verbmobil: Architecture* for the first INTARC demonstrator in April 1994. After repeatedly performing evaluation, correcting inconsistencies in the knowledge bases and the parsers, it was discovered that the standard evaluation procedure [13] led to a recognition rate that was lower than output which had been evaluated manually. For this reason a new diagnostic evaluation procedure has been developed for Linguistic Word Recognition which takes overlaps and gaps in word lattices into account. The evaluation procedure can be applied at all levels of Linguistic Word Recognition, i.e. can be used for evaluation of phoneme output, of syllable output and of word output.

# 3 Diagnostic Evaluation

Hirschman & Thompson [9] distinguish between types of evaluation in speech and natural language processing. **Adequacy evaluation** defines the fitness of a system to the task required. This they term evaluation proper. **Diagnostic evaluation** is the production of a system performance profile with respect to a taxonomisation of the space of possible inputs. **Performance evaluation** measures system performance in one or more specific areas. This type of performance serves as the basis for assessing the progress of a system. This report is concerned with diagnostic evaluation in the sense described above, with the addition that we consider diagnostic evaluation to be a more general term which covers both adequacy evaluation and performance evaluation. We define diagnostic evaluation to consist of two evaluation stages: logical evaluation and empirical evaluation. Logical evaluation is undertaken with respect to a data model. The data model defines one possible input space for the system, namely the optimal input data, and is generated top-down on the basis of labelled speech files. Optimal input data is the input a system would hope for in the ideal case. This concept is relevant only for levels of processing which explicity use structured linguistic knowledge since only here is it possible to define what optimal input would be. At the level of sentence syntax, for example, it is possible to define what the optimal input for a sentence parser would be; a single utterance which is grammatically correct. Clearly, this will not resemble real input in a spoken language recognition system but evaluating the sentence syntax level with optimal input is equivalent to verifing that all components which participate in the logical evaluation are internally consistent. Linguistic components of speech recognition systems are often criticised for assuming a near optimal input which leads to problems when the parser is coupled with an acoustic component. However, if optimal input is used for evaluation in order to develop a consistent linguistic component which has been designed also to deal with suboptimal input, then an evaluation with real data can be carried out without internal inconsistencies leading to failure. As mentioned above, logical evaluation can be applied at all linguistic levels. In particular, this report is concerned with the evaluation of the components of Linguistic Word Recognition (phoneme, syllable and word recognition).

In some ways logical evaluation is similar to testing a stochastic model with training data rather than with test data; such a procedure would not allow for participation in a competitive evaluation of the performance of several systems, but it does indicate the performance levels which can hoped to be attained on real data after tuning has taken place. Logical evaluation of linguistic word recognition components differs from the above in that a recognition rate of 100% can be achieved if the all components which participate in the evaluation are sound and complete. The stochastic model relies on a certain statistical generalisation which makes a 100% recognition rate on training data more difficult.

Empirical evaluation is defined here as evaluation on real input data. It is the recognition rate achieved by empirical evaluation which can be compared with current evaluation results in the area of speech and natural language processing.

## 3.1 Logical Evaluation using a Data Model

As mentioned above, logical evaluation of a component involves a test for soundness and completeness with respect to a data model. This notion was first presented in connection with *Verbmobil* Project 15.6 by [8].

In order to perform a logical evaluation the following steps are necessary:

- **Task**: Test all the entries in the lexicon with the grammar of the component. **Consequence**: If the grammar does not permit analysis of all lexicon entries, then either a correction of the lexicon or a revision of the grammar is required.

- **Task**: Generate optimal input data for the component using either automatically phonemically labelled data (as generated by HTK, for example) or manually corrected label files and test the component. **Consequence**: The component must at least be able to analyse what it considers to be optimal data. Otherwise the processing of the component is incorrect according to the optimal data model.

- **Task**: Generate automatically the reference files for the evaluation software using either automatically phonemically labelled data (as generated by HTK) or manually corrected label files. This must correspond to the format chosen in connection with the generation of optimal input. **Consequence**: These files serve as the basis for the evaluation and if inconsistencies are found then evaluation will not be correct.

- **Task**: Test the evaluation software for inconsistencies. **Consequence**: If the evaluation procedure is inconsistent, then a new procedure must be drafted.
- **Task**: Visualise the output lattices of the component. **Consequence**: The user can see the extent of overlap and gaps in the output lattice of the component.
- **Task**: Visualise the output of the evaluation software. **Consequence**: The user can compare this visualisation with that of the output lattices to see whether any information has been lost.

In addition, in order to avoid inconsistencies, the lexicon for the component should also be generated automatically. However, this is a knowledge acquisition task rather than a step in the logical evaluation procedure. As will be seen below, these two tasks are interrelated and important for a successful diagnostic evaluation.

As was mentioned above, the assumption is made here that in linguistic processing, a recognition rate of 100% can be achieved by the logical evaluation. Only a component which achieves this rate is sound and complete for this data model. It is not until this recognition rate has been achieved, that an empirical evaluation should be carried out.

After an iterative logical evaluation had been performed for the syllable recognition module on the ASL scenario (200 utterances of the speaker SATD), a logical recognition rate of 98.9% was achieved using the Braunschweig software. The error rate of 1.1% is due to the fact that no phonologically and morphologically relevant temporal statistics were calculated for this scenario and therefore long segments were not divided into two separate segments and therefore two overlapping syllables were generated (cf. 1). However, the visualisation indicated that both syllables had been found but that they did not stand in a connectedness relationship and therefore a substitution was assumed by the Braunschweig software. Since this type of phenonemon is likely to occur even more frequently in real data, it was decided to develop an evaluation procedure which caters for the needs of the Linguistic Word Recognition component. This evaluation procedure is term BELLE evaluation.

The BELLE evaluation procedure has been implemented and is described in section 4.7 in detail below. It is based on the notion that a reference file which only defines the connectedness relationship between units is not suffucient for evaluation purposes since signal time also plays a role. The recognised units must also correspond to a subsection of the signal. Temporal annotations for the units of reference files can be defined top-down from label files. Since it is unlikely that a recogniser of these units will recognise precisely these units at precisely these signal time points, a deviation parameter must be defined which specifies by how much the recognised unit may differ from the temporal annotations in the reference file.

The deviation parameter varies according to the size of the unit. A syllabic unit, for example, could be permitted to deviate in its endpoints up to 150 ms from the temporal annotations. Phonemic units, on the other hand, would only be allowed to deviate by say less than 30 ms. However, since the correct deviation factors for the respective levels depends on empirical testing, it must be possible to parameterise the evaluation software in order that all values can be tested. The deviation parameter which produces the best results is then the most suitable for this unit.

The new evaluation software produced a logical recognition rate of 100% for the syllable parser. Since we knew in advance that only one-segment overlaps occurred, it was possible to set the deviation parameter immediately to 30ms. This will always be the upper bound for the deviation parameter with logical evaluation using a data model. However, as will be seen in the next section, the setting of the deviation parameter plays an important role in empirical evaluation.

The software described in section 4 below is currently being used to perform logical and empirical evaluation on the Verbmobil-Scenario data.

## 3.2 Empirical Evaluation and Linguistic Word Recognition

Empirical evaluation also involves most of the steps defined under logical evaluation although the majority of the steps will have been taken in connection with logical evaluation and thus will not have to be repeated. Empirical evaluation differs to the extent that evaluation of the component is undertaken with real data rather than with optimal data which is generated top-down from labelled data. In addition to the evaluation, it is desirable to have a visualisation of both output and evaluation results. However, it is obvious that empirical evaluation is only meaningful when a complete logical evaluation using a data model with manually corrected labelled data has been performed and a recognition rate approximating 100% has been achieved.

Here also, the first set of data which was evaluated in Linguistic Word Recognition (all utterances from CD-ROM PhonDat Vol. II (1992) 'Zugauskunft' by the speaker SATD) was that agreed upon for the first INTARC demonstrator (April 1994). Before discussing the results of the empirical evaluation, a brief description of the parameterisation of the linguistic components is presented. This issue is discussed in more detail by Carson-Berndsen and Drexel in connection with the syllable recognition module in a forthcoming report.

The BELLE parsing components, the syllable parser (SILPA) and the morphoprosodic parser (MORPROPA) use the notion of parameterised linguistic models. Both parsers are parameterisable to allow for constraint relaxation and constraint enhancement. That is to say, it is possible to set the parameters so that in the case of underspecified (or unreliable) input the phonological and morphological constraints can be relaxed and phonological and morphological information an be added based on whichever information is reliable in the input. By testing the complete parameter space for each of the possible it is possible to define which units (acoustic events, phonological events) are unreliably recognised by the responsible components.

Using constraint relaxation and enhancement, an empirical evaluation of the syllable parser SILPA produced a phoneme recognition rate of 72.2% and a syllable recognition rate of 37%. As mentioned above, the phoneme recognition rate is a side-effect of syllable recognition since the phonemic units are calculated top-down from the recognised syllables. Word recognition rate was 50%.

This report is primarily concerned with the method of diagnostic evaluation for the components of Linguistic Word Recognition. More detailed results of logical and empirical evaluation of the system components will be discussed in a separate report when diagnostic evaluation has been completed for the Verbmobil-scenario data for the INTARC I.3 demonstration in April 1995.

In the remaining chapters of this report, a front-end for diagnostic evaluation of Linguistic Word Recognition is presented which is being developed at the University of Bielefeld by Verbmobil AP 15.6. The system is called BEETLE (<u>B</u>ielefeld <u>E</u>xtended <u>E</u>valuation <u>T</u>oolkit for <u>L</u>attices of <u>E</u>vents).

# 4 Bielefeld Extended Evaluation Toolkit for Lattices of Events

BEETLE is a toolkit for diagnostic evaluation in Linguistic Word Recognition which has been developed at the University of Bielefeld in connection with *Verbmobil* Project 15.6. BEETLE allows both logical and empirical evaluation to be done automatically given phonemically labelled data [1] in a predefined format, and a linguistic analysis component (currently a syllable recognition component or a word recognition component). [2]

Figure 4 shows the main BEETLE window with the possible stages of diagnostic evaluation. In addition to the steps defined in connection with logical evaluation in Section 3.1, an additional step has been incorportated which concerns the acquisition of linguistic knowledge. The lexicon for each linguistic component of word recognition is generated automatically on the basis of the phonemically labelled speech data.

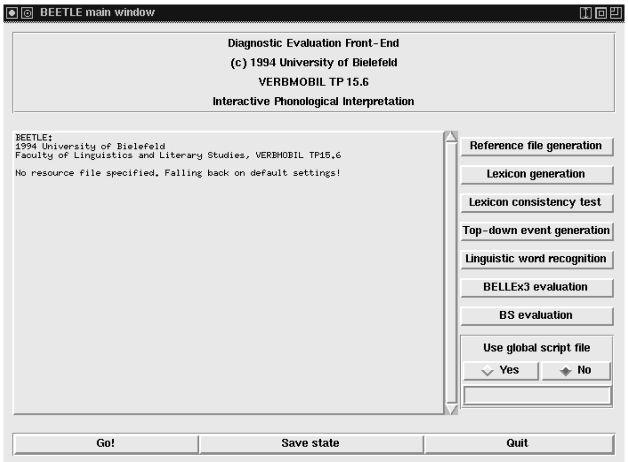


Figure 4: Beetle User Interface

In the following subsections, the individual steps in diagnostic evaluation

[1] The term phonemically labelled data as used in this report refers to either automatically phonemically labelled data or to manually corrected labelled data

[2] The main implementation of BEETLE has been undertaken by Julie Carson-Berndsen and Frederek Althoff at the University of Bielefeld. Thanks is due also to Guido Drexel, Katrin Kirchhoff, Martina Pampel, Christoph Schillo and Markus Vogt for implementation of individual tools.

are described in detail.

## 4.1 Reference File Generation

This stage of diagnostic evaluation concerns the generation of reference files for the components output unit (i.e. in this case phonemes, syllables or words). The unit is parameterised and can be set by the user. The input is a script file listing the phonemic label files in the following format:

Start-Time End-Time Label Confidence-Value

whereby Start-Time and End-Time are given in ns (/10000 to give ms).

As was discussed in Section 3.1, it was necessary to define a new evaluation procedure for Linguistic Word Recognition and therefore, two different reference files are generated: the reference file for the Braunschweig evaluation software and the reference file for the Bielefeld evaluation software (BELLE Evaluation).

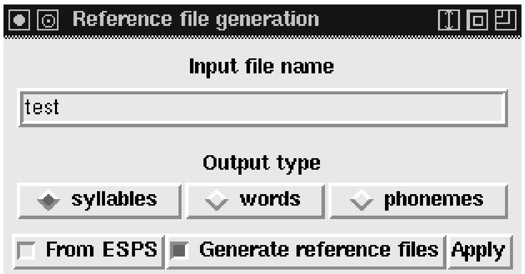


Figure 5: Reference File Generation

Reference file generation works in the following way. The phonemic label files are converted into an internal format. Then for the complete utterance all possible substrings are calculated and are presented to the user who marks the required structures using the cursor and return keys (cf. figure 6). It would be possible to insert a parser which generates only the relevant strings for the unit but this tool has been kept more general in order that other substructures such as demisyllables or bigrammes can also be marked. An example of a section of such a file is provided in figure 6 where only the relevant syllables are marked. When this marked file is stored by the user, it is converted into the label format defined above providing, in this case, a syllable label file. On the basis of this second label file, two reference files are generated.

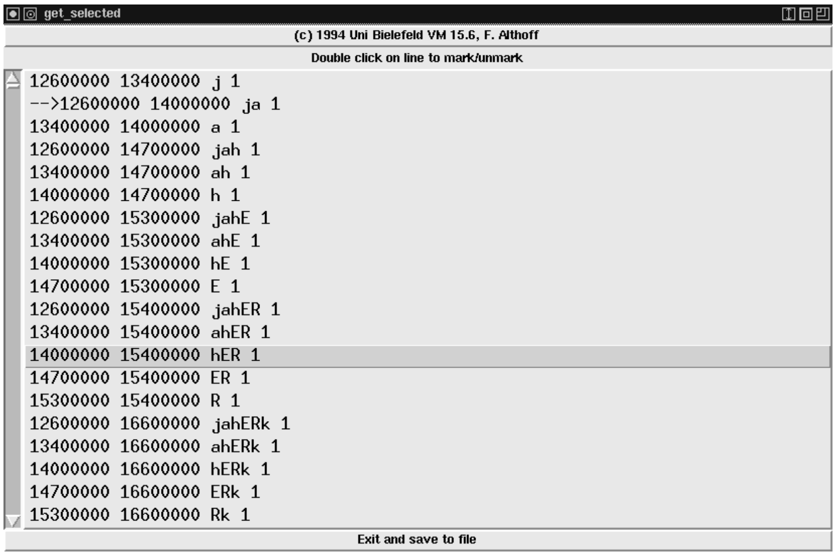


Figure 6: Marking Syllable Units for Reference Files

The first reference file is that required by the Braunschweig evaluation software. An example showing the file format is provided in figure 7. The reference file consists of an indication of the utterance number and then a list of the correct items in a defined order. The utterance number is provided at the beginning of the line due to the fact that the Braunschweig software in addition to allowing evaluation of a single utterance, also allows many utterances to be evaluated together. However, for the purposes of diagnostic evaluation, in particular in connection with visualisation of results it is necessary to be easily able to access the reference for a single utterance. For this reason, the filename also indicates the utterance number. When many utterances are to be evaluated as is described in Section 4.6 below, then the reference files may be concatenated at the time of evaluation.

```
530 IC m9C t@ fOn mYn Cn y b@ nY6n bE6k nax ham bU6k fa: r@n
531 IC bIn In k9ln Un m9Ct In 6 na hal bm StU n nax mYn Cn fa6n
532 IC braU x@ hOY t@ aI n@ f6 bI nUN ap k9ln
533 IC mYs t@ Y b6 dY sl d)6f na ham bU6k fa:n
534 IC b@ n2 tI g@ aI n@ tsuk f6 bIn dUN fOn mYn C na:x a: xn
535 gu dn mO6 N IC m9C t@ hOY d@ tsvI Sn axt Und Elf u6 a: bnts In han bU6k zaIn
```

Figure 7: Example of Reference File in Braunschweig Format

The second file format is that required by the BELLE evaluation software. In this case, the reference file consists of a set of PROLOG clauses which defined the correct unit together with its temporal annotations (start and end points) based on the original phonemic label file. The file name indicates the utterance number and therefore this is not provided in the file itself.

In addition to the generation of reference files, ESPS label files can be generated for all marked substructures of the utterance on the basis of the format defined above. A direct alignment of units to the signal is therefore possible.

## 4.2 Lexicon Generation

Lexicon generation for the components output unit (i.e. in this case again phonemes, syllables, words) can be performed according to the unit selected by the user using the intermediate notation of the reference file generation. This is the default setting. Lexicon generation expects as input a script listing files from which a lexicon is to be generated. It is possible to use files from another source rather than the default setting. These files must be in the format defined in the previous section, however. Figure 8 shows the user window for lexicon generation.

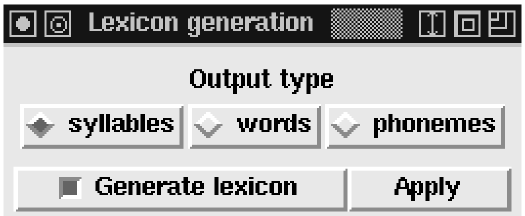


Figure 8: Lexicon Generation

The output of lexicon generation is currently component-specific. That is to say, in the context of *Verbmobil* Project 15.6, the syllable and word recognition components have their own internal lexicon formats. It is these formats which are generated here. However, the infomation contained in these lexica is feature/event-based and includes information on frequency of occurrence and average temporal duration of the units and the phonemic transcription which could clearly be output in another more general format if required.

In addition to the generation of component-specific lexica, this module is also responsible for generation the unit-specific lexicon required by the Braunschweig software. As was mentioned in Section 2, this latter lexicon is not meaningful in the context of Linguistic Word Recognition. However, it must be generated in order for the Braunschweig evaluation to function correctly.

## 4.3 Lexicon Consistency Test

Figure 9 shows the user window for the lexicon consistency test.

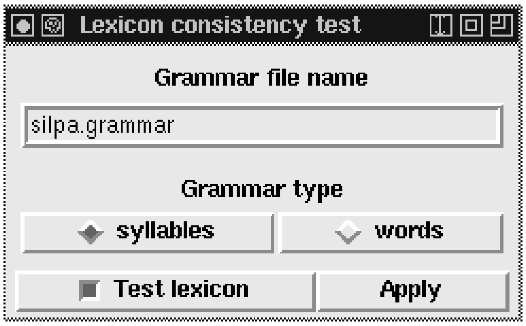


Figure 9: Lexicon Consistency Test

Lexicon consistency is component-specific and involves testing the generated lexicon (in phonemic format) with the grammar of the syllable and morphoprosodic parsers and verifying that all lexicon entries may be analysed by the grammar. The output of the lexicon consistency test is the set of lexicon entries which may not be analysed by the grammar. If this set is empty, then the lexicon consistency test is successful. As mentioned in section 3.1, this is a possible source of inconsistency of the system. If the grammar does not permit analysis of all lexicon entries, than either a correction of the lexicon or a revision of the grammar is necessary.

## 4.4 Top-down Event Generation

The tool used for top-down event generation has been described in detail in [2] and [4]. Here event structures are derived top-down from phonemically labelled data. These structures correspond to optimal input data and form the basis for the data model. Figure 10 shows the user window for top-down event generation. Input is a script of phonemic label files in the format defined in Section 4.1. The default setting is the original input script to the reference file generation. The output type is selected by the user (i.e. in this case either as input for the syllable parser SILPA or as input for the morphoprosodic parser MORPROPA). For each label in the input files, lookup is performed which maps the phoneme labels to the relevant events or features and the obligatory contour principle (smoothing) is performed (cf. [2] and [4] for further details).

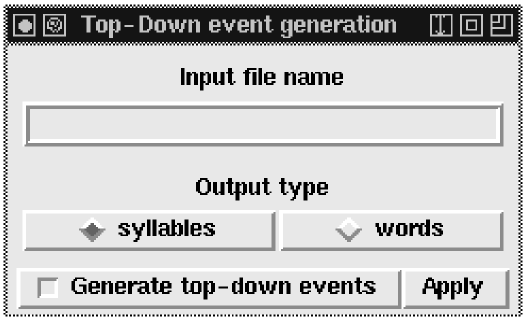


Figure 10: Top-down Event Generation

The output of top-down event generation is an optimal input file for the relevant parser which has a direct alignment to the signal.

### 4.5 Linguistic Word Recognition

This is the section which is concerned with the parser call which produces the output lattices for the evaluation. The main window is shown in figure 11. A selection is made by the user as to whether syllable parsing (SILPA) or word parsing (MORPROPA) is to be performed. Each of the parsers is parameterised (cf. section 3.2). An additional window allows the user to alter the system configuration. The default settings are shown in the figures.


Figure 11: Linguistic Word Recognition

The parameter window for the configuration of the syllable parser SILPA is shown in figure 12. The parameters are discussed by Carson-Berndsen and Drexel in a forthcoming report.

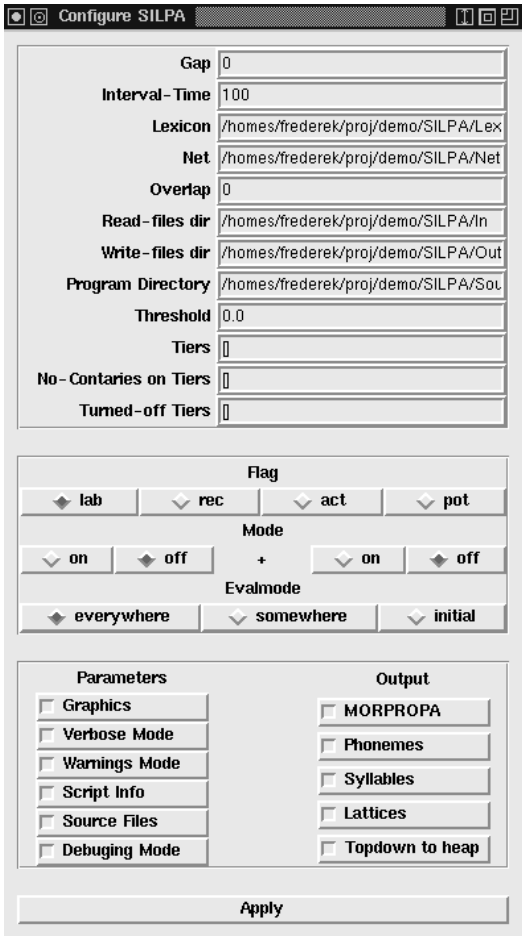


Figure 12: SILPA Configuration

Each parser component produces output lattices for their own units (i.e. syllable parser produces a syllable lattice and in addition a phoneme lattice which is derived top-down from the recognised syllables). The lattice has the following format:

Node-1 Node-2 Label Confidence Start-Time End-Time

The standard output produced by SILPA is in the following tuple notation:

⟨ Syllable—Phoneme, Start-Time, End-Time, Lex-Key ⟩

Lex-Key refers to the type of syllable recognised with respect to the lexicon. **lab** referes to those syllables which are labelled with respect to the corpus, **act** refers to those syllables which are current in the German language and **pot** refers to those syllables which are wellformed with respect to the phonotactic constraints of German (i.e. new syllables).

The configuration window for the morphoprosodic parser MORPROPA is shown in figure 13. The parameters can be set similarily to those for the syllable recogniser to allow for constraint relaxation and enhancement.

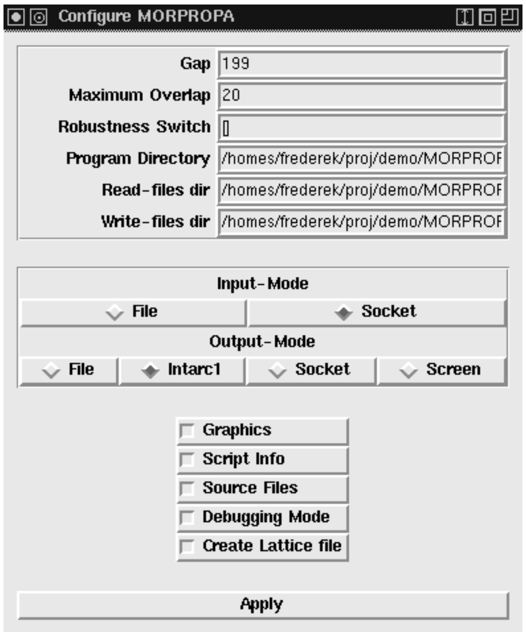


Figure 13: MORPROPA Configuration

In the next sections, two evaluation procedures for word recognition are described. The Braunschweig evaluation procedure has been described in [13]. The BELLE evaluation procedure was introduced in section 3.1 above and is discussed in more detail here.

## 4.6 Braunschweig Evaluation

Within BEETLE, the Braunschweig evaluation procedure is called to perform evaluation on the output lattices of SILPA and MORPROPA. Before this can be done, it is necessary to perform a lattice-to-chart-mapping analogously to [7] in order to guarantee at least one connected path through the lattice. In order for the Braunschweig evaluation to function within the context of Linguistic Word Recognition, a framework has been implemented in BEETLE which maps the output lattice to a chart allowing for a certain amount of overlap (as defined by the parameter *prepeval*), which calls the Braunschweig evaluation software and which optionally provides a visualisation of both the input lattice and the output chart.

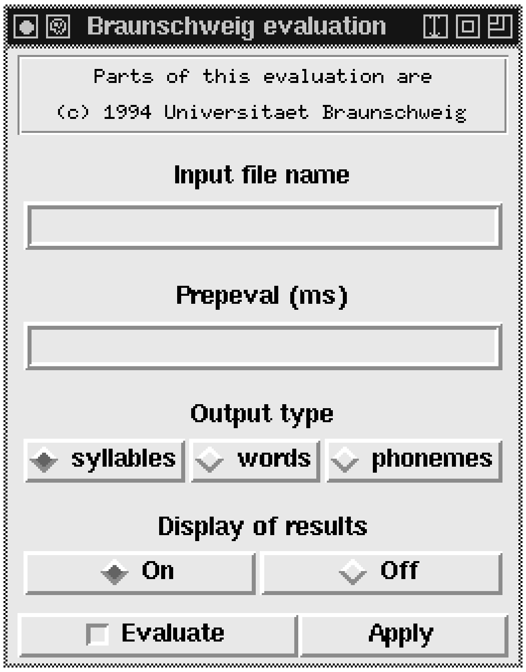


Figure 14: Braunschweig Evaluation

The framework also includes additional information on date of evaluation and on the reference files used. Examples of the text output of the Braunschweig evaluation software is given in figures 15 and 16.

```
Satzkennung:      530
Satznr:           1

REF: I C m 9 ** C t * @ f O n M Y n C N y B  @ n Y6 n ** B E6 K n
     a x H a M B U6 K  f a: R @ n
HYP: I C m 9 e: C t F @ f O n * Y n C * y e: @ n Y6 n e: N E6 M n
     a x * a * * U6 e: f a: * @ n

Fehlerstatistik

Fehleranzahl:    13
Einfuegungen:     3
Ersetzungen:      4
Loeschungen       6
```

Figure 15: Evaluation phonemes

```
Satzkennung:      530
Satznr:           1

REF: IC M9c  T@  FoN mYn cN  y ** B@ nY6n Be6K nax HAM Bu6K fa:
     R@N
HYP: IC MyNS BRE FA: mYn JA: y iN A: nY6n A:   nax X@  A:   fa:
     A:

Fehlerstatistik

Fehleranzahl:    10
Einfuegungen:     1
Ersetzungen:      9
Loeschungen       0
```

Figure 16: Evaluation syllables

The visualisation which has been incorporated into the framework around the Braunschweig evaluation was developed as a general tool for displaying hypotheses and showing their alignment to the signal. The visualisation tool, **GraphHypo** [3], displays the hypotheses with respect to either the signal times or the logical nodes of the chart. An example of the visualisation is presented in figure 18 in connection with the BELLE evaluation.

[3]implemented by Frederek Althoff, 1994

## 4.7 BELLE Evaluation

The BELLE evaluation procedure was introduced in section 3.1 above. It was developed for the evaluation of the parsing components of Linguistic Word Recognition in BELLEx3. It caters for the notion of overlaps and gaps in the hypothesis lattice and does not assume that a connected path through the lattice exists. The reference files used for the evaluation procedure take the temporal annotations of the reference units into account and a unit is regarded as recognised if, in the output lattice, there is a corresponding unit with temporal annotations which deviate from the reference unit by not more that the specified deviation parameter.

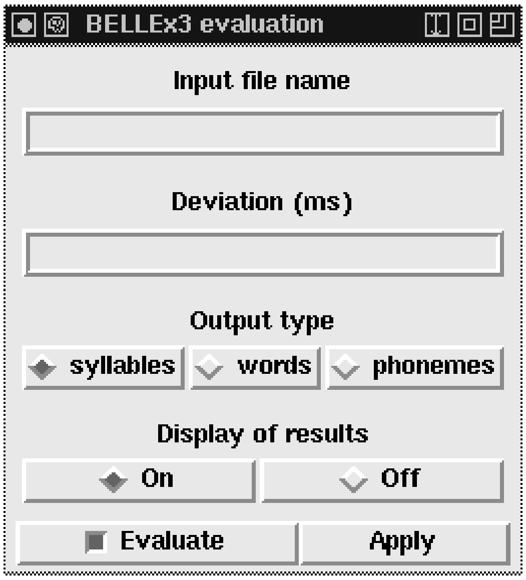


Figure 17: BELLE Evaluation

Figure 17 shows the BEELE evaluation configuration window. Input to the evaluation procedure is a script defining the files to be evaluated. The default is the script given to Reference File Generation. The deviation parameter may be set by the user, e.g. 150ms for syllable units. As mentioned above, the deviation parameter varies according to the size of the unit. Since the correct deviation factors for the respective units depends on empirical testing, the evaluation software has been parameterised in order that all values can be tested. The deviation parameter which produces the best results is then the most suitable value for this unit.

Figure 18 shows a visualisation of the output of the BELLE evaluation procedure for a syllable lattice. The shaded area represents hypotheses found in the output lattice which correspond to the reference file for this utterance. The nonshaded hypotheses represent the reference path as generated by Reference File Generation. Hypotheses which do not correspond to the reference path are not shown, for reasons of clarity. As can be seen from the figure, there is an overlap of nasality in the combination /fOnmYn/ in *von M"unchen.* These syllables would not be regarded as recognised using the Braunschweig software without an arbitrary splitting into two nasal segments. However, since the place of articulation is underspecified in the input to the syllable parser, an arbitrary splitting of such a nasal segment into two further segments is not justified.

The BELLE evaluation procedure provides a test output similar to that of the Braunschweig evaluation except it contains no reference to substitutions and insertions.

Both the Braunschweig and the BELLE approaches allow for the evaluation of single and multiple utterances. Each of the steps of BEETLE diagnostic evaluation as described in the previous subsections can be performed individually.

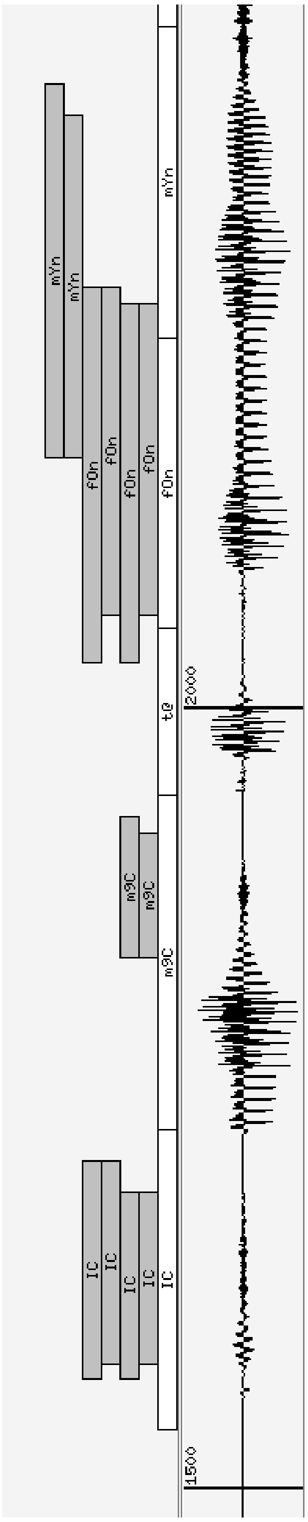


Figure 18: Visualisation

# 5 Open Issues

An issue which has not yet been considered in connection with evaluation of the components of Linguistic Word Recognition is the notion of *phonological proximity*. Standard evaluation selects the n-best paths through the output lattice with respect to the confidence values provided for each unit. Since Linguistic Word Recognition utilises underspecified phonological event structures, it implicitly defines the notion of phonological similarity between units. However, it would seem more suitable, instead of selecting the *best* hypothesis (w.r.t some confidence value) at a particular point in time, to select the hypothesis which is the nearest, phonologically speaking, to the reference unit. As was mentioned explicitly in the introductory section, it is not syllable or phoneme hypotheses which are passed from SILPA to MORPROPA but rather underspecified subsyllable events and that syllable and phoneme lattices are generated soley for the purposes of syllable evaluation. Since the syllable and phoneme hypotheses are generated by *multiplying out* the underspecified phonological event structures, the corresponding fully specified event structures are clearly more closely related phonologically than other fully specified event structures which are not subsumed by this.

In this report, there has been no discussion on evaluation criteria for Linguistic Word Recognition. This is clearly an issue which must be considered. Although it is possible for the Linguistic Word Recogniser (BELLEx3) as a whole to take part in the standard evaluation of word recognition, it has been shown in this report that such an evaluation procedure imposes restrictions which stand in opposition to the aims of interactive phonological interpretation as followed by Verbmobil AP 15.6. The aim of this report has been to show the shortcomings of the standard evaluation procedure with respect to new linguistic approaches to word recognition and to offer an alternative evaluation procedure which is in line with the notion of delayed, nonrigid segmentation. Evaluation criteria for Linguistic Word Recognition can now be drawn up on the basis of the diagnostic approach to evaluation as defined above.

# Bibliography


[1] Julie Carson-Berndsen. Ereignisstrukturen für phonologisches Parsen. ASL–TR–9–92/UBI, 1991. University of Bielefeld.

[2] Julie Carson-Berndsen. Computational tools for the development of event phonologies. In *Konvens 92 1. Konferenz "Verarbeitung natürlicher Sprache*, pages 69 – 78, Nürnberg,, 1992.

[3] Julie Carson-Berndsen. An event–based phonotactics for German. ASL–TR–29–92/UBI, 1992. University of Bielefeld.

[4] Julie Carson-Berndsen. Tools for the development of event phonologies. ASL–TR–35–92/UBI, 1992. University of Bielefeld.

[5] Julie Carson-Berndsen. *Time Map Phonology and the Projection Problem in Spoken Language Recognition.* PhD thesis, University of Bielefeld, April 1993.

[6] Julie Carson-Berndsen and Dafydd Gibbon. Event relations at the phonetics/phonology interface. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 92)*, pages 1269–1273, Nantes, 1992.

[7] L-F. Chien, K.J. Chen, and L-S. Lee. An augmented chart data structure with efficient word lattice parsing scheme in speech recognition applications. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING 90)*, pages 2: 60 – 65, Helsinki, 1990.

[8] Dafydd Gibbon. Architekturkonzepte für Worterkennung in multilingualen Dialogen. Talk presented at the Verbmobil-Workshop Architektur, Erlangen, April 1994, 1994. University of Bielefeld.

[9] Lynette Hirschman and Henry Thompson. Overview of evaluation in speech and natural language processing. Unpublished Draft, 1994.

[10] Kai Hübener and Julie Carson-Berndsen. Phoneme recognition using acoustic events. In *Proceedings of the 3rd International Conference on Spoken Languauge Processing, vol 4*, pages 1919 – 1922, Yokohama, Japan, 1994.

[11] Kai Hübener and Julie Carson-Berndsen. Phoneme Recognition Using Acoustic Events. Verbmobil Technical Report No. 15, 1994. University of Hamburg and University of Bielefeld.

[12] A. Jusek, H. Rautenstrauch, G.A. Fink, F. Kummert, G. Sagerer, J. Carson-Berndsen, and D. Gibbon. Dektektion unbekannter wörter mit hilfe phonotaktischer modelle. In *Mustererkennung 94, 16. DAGM-Symposium Wien*, Berlin, 1994. Springer–Verlag. erscheint.

[13] Michael Lehning. Ein programmsystem zur evaluierung der signalnahen spracherkennung. Verbmobil Technischer Bericht Nr. 9, 1994. Technische Universität Braunschweig.

[14] Elmar Nöth and Bernd Plannerer. Schnittstellendefinition für den Worthypothesengraphen. Verbmobil Memo 2, 1994. Universität Erlangen-Nürnberg, Technische Universität München.