A PHONETIC MORPHEME LEXICON FOR GERMAN

Andreas Mengel
Institut für maschinelle Sprachverarbeitung, University of Stuttgart,Germany

ABSTRACT

The availability of computerized lexical data is growing. In spite of this fact, little resources are available for the minimal functional units of language: morphemes. For German several morpheme lexica provide morphemes in orthographical representation, only one of them provides this information in a machine readable form. Yet, no resource is available that lists the complete morpheme inventory of German in a phonetical representation. This paper argues in favour of phonetic morpheme databases for phonetic research and speech applications. Procedures for the development of a database for German are described and first results from analyses are reported.

1. MOTIVATION

The most common conventional technique for the storage of texts is its orthographic representation. The standard minimal semantic units used in orthographic representations are words. Thus, these units are most often used as the basic entities for providing information on the correspondence between orthographical and acoustical manifestations of language: Pronunciation dictionaries represent words in both orthographical and phonetical form. The provision of access to the symbolic link between these two representations is an important factor for many applications in speech technology.

From the viewpoint of semantics, morphemes must be considered the smallest sign units. Yet, these basic semantic units of speech remain opaque for technical applications and linguistic research: For no language is there a complete database that lists the phonetical description of its morphemes. For German, many symbolic databases are available, but only few deal with morphological units ([1,2,3]) and even less - only [3], for a small subset - represent phonetic information of German morphemes.

2. WORD VS. MORPHEME

The technical value of a morpheme inventory may strongly depend on the structure of the specific language it is produced for: German is a language for which setting up a morpheme database seems very promising since German extensively uses derivational and compositional processes. German is very productive, most words can be split into two ore more morphemes.

Compared to databases which provide information on whole words and their corresponding transcriptions, a morpheme based database is easier to maintain: First of all there are less entries and any change to be applied to many entries will thus need less effort. Second, since German is so productive, a list of words for this language will never be complete and virtually never be up to date. On the other hand a word database will include many entries with very low frequencies. A morpheme database on the other hand, is only subject to little fluctuation. Thus, with a morpheme inventory at hand, a higher coverage of - German - words can be obtained.

3. PURPOSES

A phonetical morpheme database can be used both in application oriented areas and basic research. The most obvious technical applications are text-to-speech synthesis (TTS) and automatic speech recognition (ASR).

3.1 TTS

For TTS, three approaches - and their combinations - to transcription are possible: A transcription database of complete words, a morphological analysis, and a rule based letter-to-sound (LTS) algorithm. With a list of complete words and their transcriptions, new words cannot be processed, a pure LTS approach can process virtually any orthographical input, but is less error proof than a morphological system [4]. Thus any method including morphological analysis will be superior to approaches that do without.

For morpheme based transcription algorithms, mainly two methods can be distinguished. In both cases an orthographical database of morphemes and a model for their combination in words is needed. The approaches differ in the intermediate step to transcription: Either the morphemes recognized are transcribed by an LTS algorithm (transcription approach) or their phonetic transcription is provided - i.e. has been previously transcribed - in the same database and only looked up (lookup approach). The last step is equivalent again: portions of the words that could not be analyzed are transcribed by LTS rules and the transcriptions are combined to one phonetic representation of the orthographic word.

For the development of a high quality LTS algorithm, be it a self-learning approach or a set of explicit rules, a database has to be used that serves as training material (self learning method) or evaluation database (explicit rules). Thus, any LTS approach requires a transcription database and - as its output has to be tested for a finite and known set of data - its quality will only be as good as the training data set and can only be guaranteed for this limited training material.

At last, a static database will be easier to maintain than any complex rule based LTS system which will always need an exception list for those entries that are defined as extreme deviations from the standard, as it would be too costly to devise special rules for them.

3.2 ASR

In the field of ASR and for the application of HMMs on many levels it would be advantageous to add another intermediate layer between the levels of words and sounds. Adding the level of morphology would constrain the number of possible candidates and thus yield smaller error rates in speech recognition. For the training of the relevant models, the training material - orthographical words and their corresponding phonetical representations - would have to be segmented and aligned accordingly. This material could be produced in either ways described above for the morphology based approach to phonetic transcription.

morpheme category			set category	symbol	examples	
					orthographic	phonetic
prefixes	bound		closed	p	*be-, ver-, ge-*	[bə, fɛ, gə]
	free		closed	P	*vor-, an-, bei-*	[foːɐ̯, /an, baɪ̯]
lexes			open	L	*hals, tür, hof*	[halz, tyːɐ̯, hoːf]
suffixes	derivational	non-native	closed	S	*-ion, -or, -är*	[joːn, oːɐ̯, ɛːɐ̯]
		native	closed	s	*-lich, -ung, -bar*	[lɪç, ʊŋ, baːɐ̯]
	inflectional		closed	f	*-s, -t, -en, -e*	[s, t, ən, ə]

Table 1. Morpheme category, symbols used and examples.

3.3 Basic research

Pure research areas that can profit from a phonetical morpheme database are phonetics and semiotics. In semiotics, more detailed studies can be carried out that investigate functional properties of sounds and sound combinations: More easily can the forms that are preferred and sound shapes that are frequent for different morpheme categories be understood. Studies of related perceptual and articulatory processes will be carried out more reliably being founded on statistically relevant material. Across morpheme boundaries the stability and assimilation of the phonetical structure of morphemes in morpheme combinations can be studied.

4. DEVELOPMENT

In the next section details about the representation and steps of development of a morpheme lexicon are described.

4.1 Segmentation

The starting point of an analysis is the identification of the morpheme inventory. Not only stems but whole paradigms including allomorphic variants are segmented using the distributional properties of the elements as a criterion: Those smallest functional units that occur in various contexts having the same meaning or function can be called morphemes. The word entries of a data base like [3] or [5] are segmented into elements which are assigned to one of five different morpheme categories, namely: free prefixes (P), bound prefixes (p), lexes (L), non-native derivational suffixes (S), native derivational suffixes (s), and inflectional suffixes (f). For an overview of the morpheme categories, their set properties, symbols used and examples cf. table 1.

Following the segmentation approach described above, a number of words remain unsegmented. In cases where not all of the segmented material can be found in more than one context, although its phonetic structure appears morphologically complex. Such synchronous monomorphemes are also segmented if these remaining parts are equivalent to morphemes in the same context. A good example is *Himbeere* [hɪmbeːɐ̯ə] (English: *raspberry*) where *beere* (English: *berry*) can be segmented into *beer-* [beːɐ̯] and *-e* [ə]. *Him-* is left over because from a synchronous point of view, there is no other phonetic context in German in which *Him-* appears. On the other hand, this fragment is positioned in a paradigmatical relationship to morphemes like *Erd-* [/eːɐ̯d] in *Erdbeere* (English: *strawberry*). This fact is seen as a criterion to segment and include *Him-* and equivalent elements in the database. Ideally, all words are segmented into morphemes which include one no more than one vowel.

4.2 Transcription

After the segmentation of the data, the individual morphemes obtained are transcribed. The transcription method is in part phonologically oriented: Final segments that - depending on the following phonetical context - are sometimes realized voiced, sometimes voiceless, are represented by the voiced variant: *hund* (English: *dog*) is either realized as [hʊnt] in *Hund* (singular) or realized as [hʊnd] in *Hunde* (plural). The corresponding alternation is predictable by the phonetical or morphological environment. Thus, the phonetic representation of *hund* is #hʊnd #.

5. FIRST RESULTS

1,250 different polysyllabic words occurring in a speech data corpus [5] have been segmented according to the details described above [6]. They have been analyzed for their frequencies, their combinational, and their phonetical structure as described in the following paragraphs.

5.1 Frequencies

Table 2 shows the number of different types, tokens and the type/token ratio for each of the six different morpheme classes.

class	types	tokens	types/token
p	9	156	0.057
P	20	128	0.156
L	913	1,589	0.575
S	56	106	0.528
s	27	213	0.127
f	19	1,221	0.016
all	1,044	3,413	0.306

Table 2. Frequency of different types, tokens and mean type/token quotient.

The set of lexical morphemes (L) can clearly be labeled as an open class. It is more difficult to say the same for the non-native

derivational suffixes: Although their number exceeds the number of the other native morpheme categories by far, they are much less frequent than the lexical morphemes.

5.2 Information

Assuming that - because of the distributional criteria used for the identification of morphemes - morphemes within the same morpheme category have equivalent grammatical functions, the type/token quotient can be seen as a measurement of the distinctive load of the morphemes of a category. The higher the quotient, the less frequent and thus less predictable is the occurrence of one of the morphemes in a given position: Individual inflectional morphemes (f) are easier to predict than any lexical morpheme (L).

According to basic assumptions in information theory [7], the information included and transmitted by inflectional suffixes (f) would thus be least, that of the average lexical morphemes (L) would be highest.

5.3 Phonetical structure

Figure 1 represents the - mean - phonetical structure of the individual specimen of the different morpheme classes. The mean segmental complexity was measured for the initial consonant clusters (Ci), the vowels (V), and the final consonant clusters (Cf): Consonants and short vowels were counted as one, long vowels and diphthongs were counted as two elements. The sound structure bars in figure 1 are aligned to the beginning of the onset of the rhymes, i.e. the beginning of the vowels.

For table 3 the sound structures from figure 1 have been categorized. In the table a plus sign (+) is used to indicate that the substructure (initial consonants, vowels, final consonants) is rather complex, a minus sign is used to represent less complex structures. In both table 3 and figure 1, structural equivalencies of sound structures of different morpheme categories can be identified: Prefixes and lexical morphemes have more complex initial consonantal structures than suffixes; morphemes at the center have longer vowels than those at the edge; basically the same applies to final consonants.

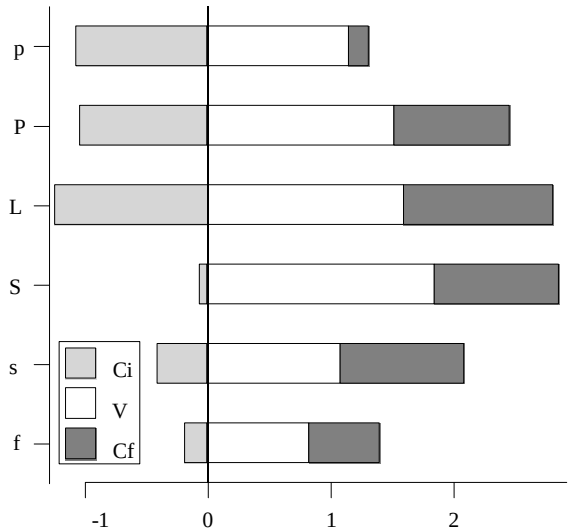

Figure 1. Mean segmental complexity of different morpheme classes.

region	p	P	L	S	s	f
initial consonants	+	+	+	-	-	-
vowels	-	+	+	+	-	-
final consonants	-	+	+	+	+	-

Table 3. Binary classification of segmental structures of morpheme classes in German. Plus signs (+) indicate complex, minus signs (-) indicate less complex sound structures.

5.4 Form and Function

From a functionalistic point of view, there should be a correlation between the information value and the form of morphemes. The less informative a morpheme, the smaller should be the effort needed to produce it. In the context of this paper and the restriction to symbolic data, the production effort of a morpheme can only be measured by its segmental complexity. It is expected that frequent and thus more predictable morphemes, that have less functional load, are less complex in their segmental structure.

Figure 2 presents the relation between the type/token ratio and the segmental complexity of the six morpheme classes defined for German. As can be seen, the assumed corespondence between form (i.e. complexity) and function (i.e. functional load) of the morpheme classes is consistent for the native morphemes analyzed in [5]: Lexical morphemes (L) which must be assumed to contribute most to the meaning of a word are most complex, inflectional morphemes (f) are least complex. The non-native derivational suffixes do not fit into the functionalistic scheme. This must be seen as an indication that their segmental structure - as they are relatively new members of the German morpheme inventory - has not yet been adapted to their functional properties. This would be another hint for categorizing them as being non-native.

5.5 Positional dependencies

Criteria for the estimation of the predictability of a morpheme are for example the predictability of the set it belongs to and the number of other members within the set. Data for these properties have be presented above. Another criterion relevant for the probability of a

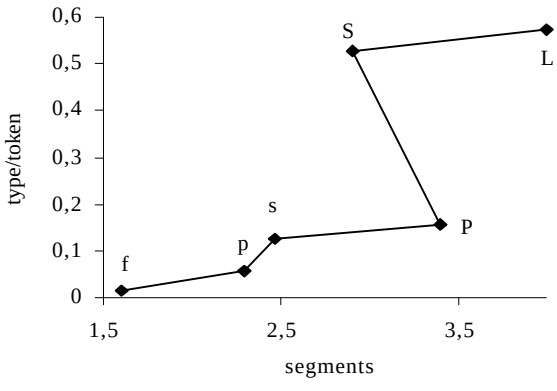

Figure 2. Functional load vs. Segmental complexity for German morphemes.

morpheme is its context. For the database segmented, the combinational probability for pairs of members of the morpheme sets have been calculated.

Table 4 presents the combinational distributions of the inventory used in [4]: The first element of a pair is defined by selecting a row (*first*), the second element by choosing one of the columns available (*second*). Thus, the probability of the combination of a lexical morpheme (L) being followed by an inflectional morpheme (f) is 19.730%. Another example is that bound prefixes (p) are most likely to be followed by lexemes (L). Percentage measurements are provided relative to all combinations. Hashmarks (#) symbolize word boundaries.

		second						
		p	P	L	S	s	f	#
first	p	-	0.199	3.495	0.022	-	-	-
	P	0.310	0.044	1.814	-	-	-	0.288
	L	0.243	0.243	4.468	1.880	4.092	19.730	3.716
	S	-	-	0.066	0.288	0.177	1.017	0.752
	s	-	-	0.088	-	0.133	1.615	2.898
	f	0.111	0.088	2.256	0.111	0.332	2.942	19.465
	#	3.052	1.880	22.185	-	-	-	

Table 4. Probabilities of pairs of members of morpheme categories.

As can be seen in table 4, the information on the left context of a morpheme can already restrict the number of possible morpheme class candidates the actual morpheme might belong to.

6. Conclusion

This paper has argued in favour of the development of a phonetic morpheme dictionary. It has discussed various advantages and suggested improvements for applications, namely speech synthesis and automatic speech recognition.

Yet, the application of a morpheme database is not limited to speech technology and phonetics only, but must also be seen as possible means of improved access to data in applications like automatic text indexing, retrieval and semantical analysis in general.

In this paper first results of the implementation of a phonetic approach to a morpheme database for German, the main criteria for the segmentation of words and the transcription of morphemes have been reported. More and larger databases are needed for many applications of this approach in fundamental research and speech technology.

The sample of data that has been used in this study to exemplify the approach. The statistical data indicate the value of the availability of a phonetic morpheme lexicon of German.

REFERENCES

[1] Augst, G. 1975. *Lexikon zur Wortbildung. Morpheminventar*. Tübingen: Narr.

[2] Kandler, G. and Winter, S. 1992-1995. *Wortanalytisches Wörterbuch. Deutscher Wortschatz nach Sinnelementen in 10 Bänden*. München: Wilhelm Fink Verlag.

[3] Baayen, R.H., Piepenbrock, R. and Gulikers, L. 1995. *The CELEX Lexical Database (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

[4] Wothke, K. 1993. Morpologically Based Automatic Phonetic Transcription. *IMB Systems Journal 32,3*: 486-551.

[5] Kohler, K.J. (ed.) 1994. Phonetisch-Akustische Datenbasis des Hochdeutschen. Kieler Arbeiten zu den PHONDAT-Projekten 1989-1992. AIPUK 26.

[6] Mengel, A. 1998. *Deutscher Wortakzent. Symbole, Signale*. Unpublished Manuscript. Technical University Berlin. Germany.

[7] Shannon, C.E. & Weaver, W. 1949. *The Mathematical Theory of Communication*. Urbana.