

A MULTI-LEVEL DESCRIPTION OF DATE EXPRESSIONS IN GERMAN TELEPHONE SPEECH

Christoph Draxler

Institute of Phonetics and Speech Communication
University of Munich
draxler@phonetik.uni-muenchen.de

ABSTRACT

The ability to efficiently and reliably process date expressions is crucial to many speech-based applications. A multi-level description which contains the syntactic deep structure, the orthographic surface form, a citation form representation, and the phonetic transcription is proposed. With this representation, high-level constraints, e.g. on syntactic structure, can be exploited to restrict the search space on the lower levels.

This multi-level representation is applied to read and spontaneous date expressions of the German SpeechDat(M)¹ telephone speech corpus, and the results of an analysis of the transcriptions of 3000 recordings are presented.

1. INTRODUCTION

1.1. SpeechDat

The European SpeechDat project aims at creating Polyphone-like telephone speech corpora for all Western European languages. The recordings are fully digital (alaw format, 8 bit quantization, 8 KHz sampling rate) on ISDN speech servers. In the first phase (SpeechDat(M), from October 1994 to February 1996), 1000 speakers were recorded in the eight languages Danish, Dutch, English, French, German, Italian, Spanish, and Swiss French [10]. Each recording consists of 39 mandatory items which are the same for all languages (digits, numbers, money amounts, time and date expressions, command words and phrases, phonetically rich sentences, spellings), and optional items (e.g. birth date, city of call, spontaneous speech) which differ from one language corpus to the other.

The German SpeechDat(M) corpus was recorded at the Phonetics institute of the University of Munich. The corpus consists of the recordings for the mandatory items plus the spontaneous speech for 500 male and 500 female speakers. Speaker ages range from 10 to 71, the dialect regions used are the 16 German states, plus Switzerland and Austria.

The present analysis of date expressions is based on one read date item, with dates ranging from 1994 to 2010², •³ 20. August 2008 or • 10. 3. 94 and two spontaneous date items: *Welches Datum haben*

wir heute? • (what date is today?) and *Nennen Sie Ihr Geburtsdatum* • (Please give your date of birth).

The recordings for the read item and today's date are on the SpeechDat(M) CD-ROMs; the date of birth was originally recorded only to serve as speaker information and thus is not on the CD-ROMs (but can be obtained separately).

All SpeechDat(M) databases will be publicly available 18 months after the end of the project, i.e. August 1997. They will be distributed via ELRA.

1.2. DATE EXPRESSIONS

Date expressions are a particularly important class of utterances in telephony applications because they relate to almost any speech-based service: telephone banking, travel information, scheduling of appointments, etc. These services are an active field of development and research (cf. *Verbmobil* in Germany [4]).

Relative date expressions indicate a date from a given (often implicit) reference point in time, e.g. *morgen*, *gestern*, *nächste Woche* (tomorrow, yesterday, next week). Absolute date expressions refer to a point in time independently of any reference, e.g. *17.4.1996*, and they are usually given in numerical format.

The numerical format is the reason why there is so little linguistic or phonological literature on absolute date expressions: there is no syntactic information in the written numerical representation found in natural language corpora; the syntactic structure is visible – and thus amenable to syntactic analysis – only in an orthographic transcription of an actual speech signal. Only recently work on the extraction of date expressions in texts has started at the Center for Information and Language processing at the University of Munich [8].

There is no phonetic literature on date expressions, because date expressions follow the usual phonotactic constraints of the language. However, for the development of speech-based applications, the ability to process date expressions efficiently and reliably is crucial, and any information that may aid in this should be made use of.

The remaining paper is structured as follows: section 2 motivates the use of multiple levels of description. Section 3 gives an introduction to date expressions in German and contains a multi-level description of the date expressions of the SpeechDat(M) database. Section 4 presents the results of an analysis of the German date expressions in SpeechDat(M).

1.SpeechDat(M) is funded by the European Commission under contract LRE-63314.

2.to get an idea of how people might utter dates in the next millennium.

3.the bullet represents the speech server's beep

2. MULTI-LEVEL DESCRIPTIONS

The "computer representation of individual languages" guidelines [3] recommend four levels of representation for speech:

- the signal itself, usually a bit stream in digital recordings,

and three symbolic representations:

- phonetic transcription, in which the signal is segmented and labelled using a phonetic alphabet, e.g. IPA symbols,
- phonological reference pronunciation or *citation form* in a language-dependent phonemic alphabet, e.g. SAM-PA
- orthographic word-based transcription, possibly augmented by markers for non-speech events.

The symbolic representation levels are mapped to each other: The transcription is performed relative to the citation form, i.e. the transcriber has the option to accept the symbol proposed by the citation form, delete it, insert a new symbol, or replace the proposed symbol by a different one [1], [3], [5], [6].

The citation form is mapped 1:1 to the orthographic representation via the lexicon which contains for each word one standard pronunciation (and possibly pronunciation variants).

In natural language processing, the basic type of data is written text, i.e. an orthographic representation. This orthography is considered as the surface form of deep syntactic structures. Even higher representation levels, e.g. semantic representations, are based on the syntactic structures.

Grammar rules describe the organization of words and syntactic substructures into larger syntactic structures; these rules can be used top-down for the generation, or bottom-up for the analysis of surface forms. Grammar rules may take arguments which restrict their applications to specific contexts; thus, the grammar can be held modular, and subgrammars can be used for the analysis of restricted sublanguages.

A multi-level description that contains the levels outlined above is useful because higher-level syntactical constraints strongly restrict the variation of items on the lower levels, and thus can restrict the search space. This is particularly important in languages with a rich case system, e.g. German. Furthermore, multi-level descriptions have shown to be a clear and flexible way of representing data on different levels of representation [1].

In this paper, three levels of representation will be used for the description of date expressions in German:

- citation form
- orthographic representation
- syntactic structure with arguments.

A segmentation and labelling of the speech data is currently under way, but was not yet available for the present analysis.

3. DATE EXPRESSIONS IN GERMAN

In transcriptions of date utterances in German, date expressions take on a large number of surface forms:

1. *der siebte elfte neunzehn hundert fünfundneunzig*
2. *den elften Oktober neunzehn hundert
fünfundneunzig*
3. *am zwölften vierten vierzig*
4. *siebzehnter zehnter dreiundsechzig*
5. *dreißig null neun siebenundvierzig*

To keep the dictionary small, numbers larger than twenty are written as sequences of isolated words instead of one long word as would be normal in German.

3.1. Vocabulary

der, *des*, and *den* are determiners. *am*, *vom*, *zum*, and *nach* are prepositions, with *am* and *vom* short forms of *an dem* and *von dem* respectively; *nach* requires *dem* to follow.

The case of a date expression depends on the determiner:

Determiner	Case
der	nominative
des	genitive
dem	dative
den	acusative

Table 1: Determiners and Case

The format can be either ordinal or cardinal. If a determiner is present, the format is ordinal (cf. examples 1, 2, 3); otherwise, both ordinal and cardinal formats are possible (cf. examples 4, 5).

For days, the vocabulary consists of ordinal numbers from 1 to 31, and of the cardinal numbers from 0 to 31. 0, pronounced *null*, is used to disambiguate sequences of cardinal numbers, e.g. *eins eins zwei fünfundneunzig* could be 11.2.95 or 1.12.95; *null eins eins zwei fünfundneunzig* can only be 01.12.95. In speech, the variant *zwo* is commonly used for *zwei* – with the same distribution as *zwei*, i.e. in ordinal or cardinal format, in isolation or in compound number words. The reason for this variant is that *zwei* and *drei* are hard to distinguish under noisy conditions, e.g. on the telephone. Furthermore, a variant *siebenter* of the ordinal *siebter* is also quite common, especially in hyper-clear articulation.

For months, ordinals and cardinals from 1 to 12 are used, or the month name. In speech, *Juno* and *Julei* are variants of *Juni* and *Juli*, which again are difficult to distinguish in the presence of noise.

The year vocabulary consists of cardinals from 1 to 19 for the centuries with the exception of 10, cardinals 1 and 2 for millennia, and cardinals from 1 to 99 for the year within a century, and the words *hundert*, *tausend*, and *und*.

3.2. Syntactic structures

The grammar rules for the syntactic structures are written in the Definite Clause Grammar (DCG) formalism provided by the Prolog programming language [2].

DCG rules consist of a head, the symbol `-->`, and a body which is a conjunction or disjunction of goals. Parameters are used to pass information between rules; upper case arguments are logical variables, lower case arguments are constants.

The head and the goals in the body are nonterminals. Terminals are written in square brackets, and goals in curly braces are regular Prolog goals:

```
date_expression(date(Det,Date),Case,Format) -->
    junk,
    determiner(Det,Case,Format),
    date(Date,Case,Format).

date_expression(date(Date),nominative,ordinal) -->
    junk,
    date(Date,nominative,ordinal).

date_expression(date(Date),acusative,ordinal) -->
    junk,
    date(Date,acusative,ordinal).

date_expression(date(Date),Case,cardinal) -->
    junk,
    date(Date,Case,cardinal).

determiner(det(Det),Case,Format) -->
    [Det],
    {determiner(Det,Case,Format)}.

date(date(Day,Month,Year),Case,Format) -->
    day(Day,Case,Format),
    month(Month,Case,Format),
    year(Year).

date(date(Day,Month),Case,Format) -->
    day(Day,Case,Format),
    month(Month,Case,Format).

day(day(Day),Case,ordinal) -->
    ordinal(Day,Case,1,31).

day(day(Day),none,cardinal) -->
    cardinal(Day,1,31).

month(month(Month),Case,ordinal) -->
    ordinal(Month,Case,1,12).

month(month(Month),_,cardinal) -->
    cardinal(Month,1,12).

month(month(Month),_,_) -->
    month_name(Month).

year(year(neunzehn,hundert,Year)) -->
    [neunzehn,hundert],
```

```
([und] ; []),
    cardinal(Year,1,99).

year(year(neunzehn,Year)) -->
    [neunzehn],
    cardinal(Year,1,99).

year(year(Year)) -->
    cardinal(Year,1,99).

year(year(Cardinal,tausend,Year)) -->
    cardinal(Cardinal,2,2),
    [tausend],
    ([und] ; []),
    cardinal(Year,0,50).

year(year(Cardinal,tausend)) -->
    cardinal(Cardinal,2,2),
    [tausend].
```

The case and format information is stored in the lexicon, e.g.

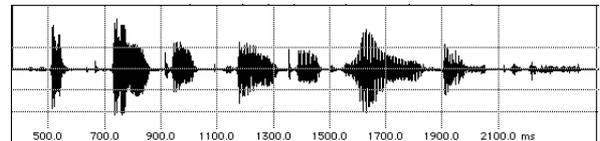
```
lex(erster,'q E 6 s t 6', ordinal,1,[nominative]).
```

This grammar uses the parameters to a) build a syntax structure of the transcription, and b) propagate case and format constraints to the right: once a marker for a case or format is found, the rest of the utterance has to be of the same case.

3.3. Example.

The signal file A00000D0 (A603S01.WAV on the proceedings CD-ROM) is described as follows:

Signal



Phonetic segmentation

```
433.9    #word_begin
433.9    z
512.8    i: - I
569.5    p
641.9    t
...
```

Citation form

```
z i : p t s e : n t @ r
t s e : n t @ r
d r a I U n t z E C t s I C
```

Orthography

```
siebzehnter zehnter dreiundsechzig
```

Syntax structure

```
Date = date_exp(date(day(siebzehnter),
month(zehnter), year(dreiundsechzig))),
Case = nominative,
Format = ordinal
```

4. RESULTS

Table 2 shows the length distribution for the three different types of date expressions with the noise markers in the transcriptions removed (markers for truncated signals and word fragments are counted).

Word count	Date of birth	Today's date	Prompted date
0	2	10	5
1...2	22	24	12
3...4	432	338	508
5...6	499	527	458
7...8	41	76	15
9...	4	24	2

Table 2: Length of date expressions with noise markers removed

85.8% (=2573) transcriptions could be parsed by the grammar. 90.7% of the birth date transcriptions could be parsed correctly, vs. 89.3% for today's date and 77.3% for the read date expressions. Note that the grammar is strict: words (including hesitation markers, word fragment, and other markers) not belonging to either the day, the month, or the year vocabulary may appear only before the beginning of the date expression proper.

97.9% of the date expressions that could be parsed were ordinal vs. 2.1% (=53) cardinal. Within the cardinal date expressions, the date of birth occurred 45 times, today's date 7 times, and the prompted date once. One reason for the rather high frequency of the cardinal format for the date of birth could be that this date is often uttered in situations that require stereotypical responses, e.g. when filling out forms.

6.3% (=163) of the date expressions were preceded by a determiner. 150 of these determiners occur in the transcriptions of today's date; the question used to elicit this date is most naturally answered by an elliptical date expression in acusative format.

35% of the parseable spontaneous date expressions contained month names, but there is a great difference between the two date expressions: 40.1% of today's date contained a month name, compared to only 28.9% in the birth date.

16.5% of the date expressions could not be parsed. The main reason for this are word fragments due to truncated signals or repair phenomena by the speaker. Word fragments cannot be found in the dictionary, causing the parser to fail. With both truncated signals and transcriptions with repairs removed (which make up 10.9% of all transcriptions), 3.6% of the transcriptions cannot be parsed. In these cases,

- at least one of the mandatory date items day or month is missing,
- the transcription contains extra items inside the date expression proper, e.g. hesitations,

- the transcription is incorrect,
- the transcriptions is empty, or
- the speaker did not produce the expected utterance.

The degree of correctness of a transcription depends largely on the transcription conventions. In SpeechDat, a minimal transcription inventory was defined, which contained the orthographic form of words as given by a reference dictionary (the Duden for German), and markers for speaker and non-speaker noise, word fragments, and truncated signals. It is clear that for example dialectal variations cannot be represented adequately in the standard orthography.

The results obtained from the SpeechDat(M) corpus represent only a fraction of the possible date expressions in German. In SpeechDat II it must be taken care that speakers are prompted to provide date expressions in all possible cases and formats.

Despite these limitations, the multi-level description of speech signal transcriptions has provided some insights into date expressions in speech, and has provided frequency data for the major classes of date expressions.

5. References

1. Atmanspacher, S., Burger, S., Draxler, C., Kipp, A., Scheer, C., Schiel, F., Wesenick, M.-B. (1995), Partiturformat für die Darstellung unterschiedlicher Repräsentationsebenen von gesprochener Sprache, *Verbmobil Memo 90*, September 1995
2. Clocksin, W. F., Mellish, C. S., *Programming in Prolog*, Springer Verlag, Berlin, 1981
3. Draxler, C. (1995), Introduction to the *Verbmobil-PhonDat Database of Spoken German*, Prolog Applications Conference PAP 95, Paris
4. IPA. (1989): The IPA Kiel Convention Workgroup 9 report: Computer coding of IPA symbols and computer representation of individual languages. *Journal of the International Phonetic Association* **19**, 81-82.
5. Karger, R., Wahlster, W. (1994), *VERBMOBIL Handbuch*, *Verbmobil Technisches Dokument 17*, Saarbrücken: DFKI
6. Kohler, K.J., Lex, G., Pätzold, M., Scheffers, M., Simpson, A.P., Thon, W. (1994), *Handbuch zur Datenaufnahme und Transliteration in TP 14 von VERBMOBIL - 3.0*, *Verbmobil Technisches Dokument 11*, Kiel: IPDS
7. Pompino-Marschall, B., (1995) *PhonDat -Verbundvorhaben zum Aufbau einer Sprachsignaldatenbank*. FIPKM 30, 1992
8. Thalmeier, Karin: private communication
9. Winski, R., Senia, F., Conner, P., Hüb-Umbach, R., Constantinescu, A., Niedermair, G., Moreno, A., Trancoso, I., (1995): Specifications of Telephone Speech Data, Deliverable 1.4.1, SpeechDat internal report