The TASX-environment: an XML-based toolset for time aligned speech corpora

Jan-Torsten Milde & Ulrike Gut

Department of Linguistics and Literary Studies
University of Bielefeld, Germany
milde@coli.uni-bielefeld.de, gut@spectrum.uni-bielefeld.de

Abstract

This paper describes the design and implementation of an XML-based corpus environment for multi-tier annotated speech data. The TASX-environment (TASX: Time Aligned Signal data eXchange format) constitutes the technical basis for a corpus designed to explore the acquisition of prosody by second language learners. It supports all aspects of the corpus setup procedure: XML-based annotation of the speech data, all transformation of non XML-annotations, and the web-based analysis and dissemination of the data.

1. Introduction

In this paper we describe ongoing research in the design and implementation of an XML-based corpus environment for complex annotated speech data. The development of the corpus environment is part of the LeaP project, which explores the acquisition of prosody by both second language learners of German and English. In a period of two years a large set of audio and video recordings of second language learners' speech will be made and phonologically annotated. From this data an XML-annotated spoken language corpus will be set up. The model is based on a client/server approach. For performance reasons the XML-annotated data can be stored in a relational database. The XSL-T-based transformation of the data is a server sided process. The TASX-environment presented here supports the complete corpus setup procedure: XML-based annotation of raw speech data, the transformation of non XML-data and the analysis and dissemination of the corpus.

The paper is organized in five sections. First, a short overview of the LeaP project will be given, which explains the specific requirements for the TASX-environment. In the next section the underlying XML-based TASX format will be explained and the components of the TASX-environment will be described in more detail. In section 4, we will then explain how the LeaP corpus has been set up with the TASX-environment and how the data can be linguistically analysed. Finally, a short conslusion will be given.

2. The LeaP project

The LeaP (Learning Prosody) project[1] explores the acquisition of prosody by second language learners of both German and English. It focuses on three areas of prosody: stress assignment on both the word and the phrase level, sentence intonation and speech rhythm. So far, the acquisition of second language prosody has not attracted a large amount of research but it has nevertheless often been proclaimed to be nearly impossible (Boyle, 1987). However, this assumption of non-attainment so far has only been supported by a few single case studies (Archibald, 1998).

In addition, the form and function of gestures in non-native speech are analysed. It is hypotheized that transfer and interference from the native language as well as an adapted variability and frequency of gestures will occur. The alignment of gestures with prosodic features will be explored.

Research in the LeaP project is based on a large corpus of second language learners' speech. The focus lies on two main research questions: first, a detailed description of the non-native prosody within the latest theoretical frameworks and a comparison to native speakers' prosody will be carried out. It is assumed that second language prosody constitutes a good testing ground for theoretical concepts in prosody and might provide evidence for their further development. Equally, non-native gestures are described and compered to native gestures.

The second line of research is concerned with the question of whether and how prosodic aspects of a foreign language can be learned. The project thus investigates the extralinguistic factors such as personal variables (e.g. native language, age at the beginning of language learning, motivation, musicality) and the type of teaching method that might enhance the outcome and speed of the acquisition process. The LeaP experimental design comprises three treatment groups, who undergo intensive prosodic training in English and German of up to one and a half years duration, as well as longitudinal studies of language learners in natural exposure settings.

For both research questions a multitude of data of various types are being collected: the corpus of spoken language will consist of at least 400 recordings of between 2 and 10 minutes length. It comprises three different speech styles: read speech, prepared speech (a retelling of a story) and free speech. In the first ten months of the project, 184 audio and two video recordings have already been made. In addition to this speech material, meta data have been collected for every speaker. This consists of personal data such as the learners' age, sex, native language and the onset of learning of the second language, as well as ratings of motivation and interest.

The annotation of the speech material is carried out using ESPS/waves+ with six different tiers (see figure 1). On the first tier, the phrase tier, phrases are annotated as well as the events occurring between them (e.g. pauses, laughter, noise). On the second tier, the word tier, each word is transcribed orthographically. On the third tier, the syllable tier, each syllable is transcribed in SAMPA. On the fourth tier, the rhythm tier, the vocalic and consonantal parts of

[1] http://www.spectrum.uni-bielefeld.de/LeaP/

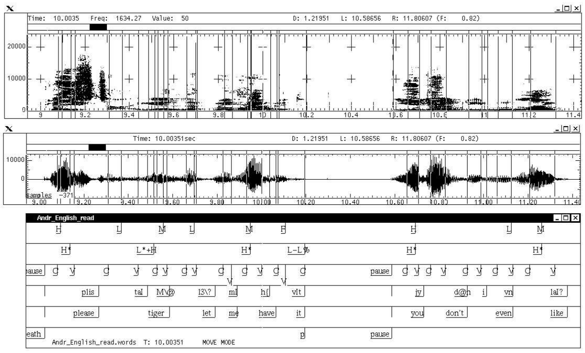
Time: 10.0035 Freq: 1634.27 Value: 50
D: 1.21951 L: 10.58656 R: 11.80607 (F: 0.82)
Time: 10.00351sec
Andr_English_read
H L M L M F H L M
H* L*+H H* L-L% H* H*
pause C V C V C V C V C V C C V C V C pause C V C V C V C V C V
plis taI MV@ l3\? mI h{ vIt jy d@n i vn laI?
please tiger let me have it you don't even like
eath p pause
Andr_English_read.words T: 10.00351 MOVE MODE

Figure 1: Prosodic annotation in the LeaP project is carried out on six tiers: a phrase, an orthographic tier, a syllable tier, a rhythm tier, a tone tier and a pitch tier.

the speech are annotated. On the fifth tier, a transcription of intonation in a ToBI (Silverman et al., 1992) style is carried out. The sixth tier, the pitch tier, contains an annotation of highs and lows in the pitch contour. This means that for each recording there are approximately 3000 time stamps. In the first ten months of the project, 129 recordings have been annotated in this fashion.

3. The TASX format

A central aspect of our research is to explore up to which point current standard XML technology (XML, XSL-T, XSL-FO, XPATH, SVG, XQUERY) can be used to model linguistic databases, to transform, query and distribute the content of such databases and to perform adequate linguistic analysis. As a result, all linguistic data in our system is stored in an XML-based format called TASX: the Time Aligned Signal data eXchange format.

A TASX-annotated corpus consists of a set of sessions, each one holding an arbitrary number of descriptive tiers, called layers. Each layer consists of a set of separated events. Each event stores some textual information (e.g. a syllable) and is linked to the primary audio data by two time stamps. Relations between events on different tiers can be encoded by defining links using the ID/IDREFS mechanism of XML. Finally, arbitrary meta-data can be assigned to the complete corpus, each session, each layer and each event. It might be sensible to extend the meta data description in a way that tree structured data can immediately be described by XML annotations. Currently we rather use the simpler version with linear structure. The following DTD fragment formalizes the TASX format:

```
<!-- corpus data -->
<!ELEMENT tasx (meta*,session+)>

<!ELEMENT session (meta*,layer+)>
<!ELEMENT layer (meta*,event+)>
<!ELEMENT event (#PCDATA,meta*)>

<!-- meta data -->
<!ELEMENT meta (desc*)>
<!ELEMENT desc (name,val)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT val (#PCDATA)>
```

```
<!-- atributes -->
<!ATTLIST session
        s-id CDATA #REQUIRED
       day CDATA #REQUIRED
        ref IDREF #IMPLIED
        month CDATA #REQUIRED
        year CDATA #REQUIRED>

<!ATTLIST layer
       l-id CDATA #REQUIRED
       ref IDREF #IMPLIED>

<!ATTLIST event
          e-id CDATA #REQUIRED
          start CDATA #REQUIRED
          end CDATA #REQUIRED
          ref IDREF #IMPLIED
          mid CDATA #IMPLIED
          len CDATA #IMPLIED>

<!ATTLIST meta
        m-id CDATA #REQUIRED
        ref IDREF #IMPLIED
        access CDATA #IMPLIED
        level CDATA #IMPLIED>
```

Despite ofs it simplicity, the TASX-format is powerful enough to encode most of the corpus annotation formats currently in use. Indeed a number of format transformation programms have been implemented. For example, in order to reconstruct the equivalent annotation graphs (Bird and Liberman, 1999) representation of a TASX annotated corpus, one only has to collect the time stamps encoded in the start and end attributes of the event tags, sort them and then produce the timeline. Finally the time stamps of the events have to be replaced by references to the timeline.

3.1. The TASX-annotator and the corpus engine

The complete TASX-environment consists of:

- tools for the annotation of empirical language data (video and audio material),
- an input mask for processing meta data
- programs for the transformation of various formats of linguistic standard software (Transcriber, Praat, ESPS/waves+, SyncWriter, Exmaralda etc.)
- a set of programs for linguistic analysis of the TASX-annotated data, and
- a corpus system for the distribution of language data via the internet, including interactive corpus query and multimodal data display in a standard web browser.

In the following sections these modules will be described in more detail (see also (Milde and Gut, 2001)).

3.2. The TASX-annotator

The TASX-annotator is a central component of the TASX-environment. The tool allows the annotation and transcription of video (multi-channel) and audio data (see figure 2).

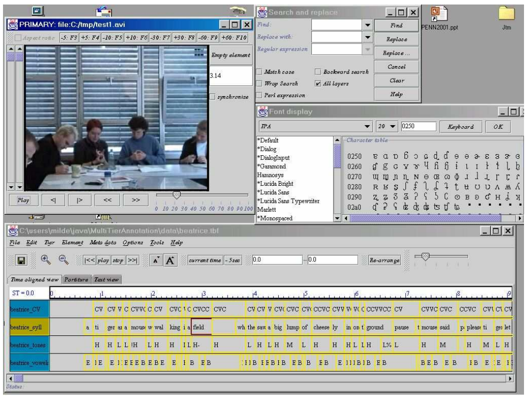

Figure 2: A screenshot of the TASX-annotator. In the bottom half, the main panel is visible, where the time aligned tier view has been selected. On top of the main window, the font selection panel is visible (showing some IPA characters) and above it the find tool. In the upper left corner the video display can been seen.

The programm is very user friendly and can be used without a high level of computer skills. It is possible to completely control the tool by either mouse *or* by keyboard shortcuts. Video and audio playback can be controlled by a foot switch. Different data views are programmed (time-aligned partiture, word-aligned partiture, sequential text view) to make annotation as effective as possible.

The time aligned view is organized as a two dimensinal grid of infinite size. A layer is presented as a horizontal tier of events. The order of the layers is arbitrary and can be changed instantly. The user is able to define time intervals by dragging the mouse. Each time interval represents an event. The event is displayed as a graphical box which can be selected and moved with the mouse. The content of an event is entered in an additional text field. Any (unicode) font (e.g. IPA fonts, HamNoSys fonts etc.) available for the operating system can be used for the transcription. The user can choose font and fontpage from a table displaying all characters of the selected font. It is also possible to define a virtual keyboard which maps the given keystrokes to arbitrary characters of the target font.

A separate video playback window will opened up for each video file making it possible to e.g. display multiple perspectives of the same scene. The video playback is synchronized with the transcription. For audio transcriptions an oszillogram is calculated and is displayed inside the main window.

In the text view the data can be manipulated in a standard text editor panel. The content of the editor represents the layer and each line represents an event. A list selection box allows switching between different layers. It is possible to transfer text from standard text editors, e.g. Microsoft Word, by cut and paste operations. In order to additionally speed up the transcription process, a word completion function has been implemented for the text view. Entering the initial letter of a word and consecutively pressing CTRL+L will bring up all words starting with this letter. Once the text is tranferred into the TASX-annotator, the events still have to be aligned with the primary audio and video data. Switching back to the time aligned view and moving the events with the mouse makes this task quite simple.

In the partiture view the data cannot be edited. In practice this means that the data is transformed into an HTML table and then displayed to the user. A number of different HTML formatted views have been designed. The views can also be saved to external files and loaded into standard web browsers.

One potential strength of the TASX-annotator is its manner of handling the export/import of XML based information. A standard way of solving this problem would be the implementation of a set of format specific XML parsers which construct the internal representation (e.g. JDom) of the XML file. While powerful integrated development systems such as *Sun's Forte for Java* make the design of such XML handlers simpler, it still remains a complex task to implement such a parser. In the TASX-annotator we follow a different approach. The system integrates an XSL-T processor (saxon), making it easy to perform on the fly data transformations. The import of an XML-file is split into two steps: first an XSL-T stylesheet transforms the XML file into TASX, second another XSL-T stylesheet will transform the TASX file into a simple text oriented format. This format can be loaded efficiently.

3.3. Transcoding tools

The development of tools for the TASX-environment is based on the concept that a re-implementation of functionalities already available in other speech processing software is not necessary. Established speech software such as Praat or ESPS/waves+ do not need to be duplicated.

TASX	import	export
Annotation graphs	XSL-T	XSL-T
Exmaralda	XSL-T/Java	Java/XSL-T
HTML-table	–	XSL-T
HTML-partiture	–	XSL-T
RTF	–	XSL-T/Java
Anvil	XSL-T	–
Praat-label	Perl/XSL-T	XSL-T
ESPS-label	Perl	XSL-T
ESPS-freq	Perl/XSL-T/Java	XSL-T
SyncWriter	Perl	–

Table 1: List of currently implemented transcoding tools. The table shows the programming languages used to implement the transcoders.

The TASX-environment therefore focuses only on the development of transcoding filters from and into various formats. These include: Praat/freq, Praat/label, ESPS/waves+, ESPS/F0-analysis, Transcriber, annotation graphs stored in XML, SyncWriter and basic text formats (see table 1). In addition, filters for data import and export of the Exmaralda system (Schmidt, 2001) are available. Most of these components are implemented in Java, transformations are defined in XSL-T and a smaller number of additional tools is written in Perl (mainly to transform non-XML data).

3.4. Pause tracker

To speed up the annotation process a pause tracking programm has been developed. The programm separates speech from pauses and generates a TASX annotated XML document with two tiers, one holding all pause events, the other one holding all speech events.

The tracker uses Praat (Boersma, 2001) to perform the actual speech analysis. It simply calculates the pitch curve of the audio signal. If no pitch is detected, then non-speech is assumed, otherwise speech. In a second step, the results of this classification are combined to continuous stretches of pauses/speech. Finally the TASX conformant output is generated.

The pause tracker has shown to work quite reliably on a set of recording in different languages (Japanese, English, German, Saterfriesisch, French, Ega). Even if tracking is far from perfect, the transcriber gets a good presegmentation of the signal. This allows to move very quickly through the file, possibly performing minor adjustments to the boundaries or combining a set of separated events of one speaker.

While the pause tracker gives good results when doing conversational analysis it is not of much help for fine grained phonetic research. Here a tracking system for vowels and consonants would be very useful. Garcia et.al. are working on such a system (Garcia et al., 2002)

3.5. The corpus system

The main function of the corpus system constitutes the internet-based dissemination of the corpus data. With the currently implemented interface it is also possible to inspect and query the speech corpus, to listen to the audio material and to display the graphic representation of the waveforms in a standard web browser. We make use of the built-in features of the web browser here. Furthermore, the PAX-tools (Gibbon and Trippel, 2001) for displaying the intonation contour, the intensity and the spectrogram of the selected regions in the audio file can be integrated.

When playing back the sound file, both the audio parts and the waveform images are generated automatically by a small Java servlet program. The servlet parses the XML-annotated corpus, extracts the time stamps of the relevant events and then cuts out the corresponding parts of the original sound file.

The corpus system is split into two larger subcomponents: the *information pool* and the *corpus engine* (see figure 3). The information pool stores the primary data (raw audio data) as well as the XML-annotated transcriptions of the audio files. The corpus engine consists of five subsystems:

1. Web-client: the interactive user interface is completly defined to run in a standard web browser. We are using HTML-query forms which activate services on the server side to generate XSL-T-filters processing the data. Waveforms are displayed using SVG. This will allow the user to select parts of the sound signal and to perfom more complex phonetic analyses.

2. Web-server: the web server distributes the corpus information in several standard formats (XML, HTML, PDF, SVG, WAV).

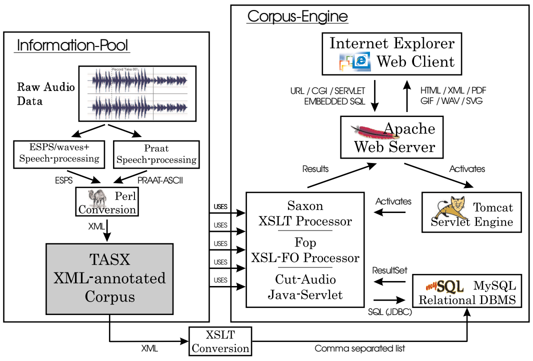

Figure 3: The system architecture of the corpus system. The corpus system is split into two subsystems: the information pool (left) storing the TASX-annotated data and the corpus engine (right) distributing the data over the internet.

3. Servlet-engine: the servlet engine activates the suitable services on the server side (transformation of XML-annotated data, on-the-fly phonetic analysis, generation of graphics).

4. Servlets: a set of TASX/XML-aware servlets are used to transform the data in numerous ways: generating HTML to be displayed in the browser, generating PDF to be printed out, generating wavefiles and images of the waveforms. XSL-T and XSL-FO are used to perfom the transformations. The servlets have access to the information pool and the relational database.

5. Relational database: in order to improve the system performance, the XML-annotated corpus data is stored in a relational database. The database basically replaces a standard file system. An XSL-T-program translates the XML-annotated corpus data into a suitable format for the DBMS.

The implementation of the corpus system is based on open source software. The TASX-annotator is a pure Java application; all other tools are smaller XSL-T and perl scripts. As a result, the complete TASX-environment runs on Windows and Unix platforms. The software will be distributed under GPL and can be downloaded from our website[2].

3.6. Statistical analysis

In the inital design phase of the TASX system we planned to implement the statistical analysis in XSL-T and Java. Indeed, a number of smaller programs have been realized in this technique. Unfortuneatly it quickly became evident, that XSL-T is not suited to perform such calculations on larger sets of data. It lacks high precision arithmetic functions and consumes to much memory. When using external Java functions, a large number of data conversions have to take place. Also the resulting code is very hard to read and debug.

Instead we have chosen to use the R system, an open source implementation of the `S-Plus` statistics language

[2] http://coli.lili.uni-bielefeld.de/∼milde/tasx/

(Ihaka and Gentleman, 1996), (Venables and Ripley, 1999). R implements all major statistical tests and calculations and is equipped with a large number of high level graphic routines to generate visually informative presentations of the results. Even more important it includes efficient input/output routines to load and save semistructured data (either XML-annotated or plain ascii text).

4. Setting up the TASX-annotated corpus

Once the prosodic annotation as described in section 2. is completed the TASX-annotated corpus can be set up.

4.1. LeaP data conversion

First , the manually labeled ESPS/waves+ files are being converted into the XML-based TASX format. This is done automatically by a small perl program called `esps2tasx`. The converter is able to take a whole set of ESPS/waves+ files and transform them into one large TASX-annotated corpus. For each of the speakers a set of additional information has been collected. This information will be stored as session meta data and will be added to the TASX-annotated corpus.

The TASX-annotated data is finally stored in a relational database. This is done to improve the performance of the corpus system. Each session is stored as a binary large object.
After the data has been transformed into the TASX-annotated form, it becomes possible to use the complete set of tools of the TASX-environment.

4.2. Analysis of the prosodic data

The phonetic analysis of the LeaP data is carried out in a semi-automatic style, supported by various TASX analysis tools. For the calculation of the speech rhythm according to Ramus et al. (Ramus et al., 1999) for example, the information of the fourth tier as described in section 2. is taken. The length of all vocalic (V) and all consonantal (C) parts of the utterances in the recording are calculated and their standard deviation (ΔV and ΔC) is computed, as well as the proportion of vocalic intervals (%V) across the entire recording. These measurements have proved useful for the description of the difference in speech rhythm between languages and varieties of languages (Gut and Milde, 2002). The results for all speakers are illustrated in an automatically generated graph (see figure 4).

Similarly, the analysis of the speakers' pitch range is carried out semi-automatically. From the time stamps of the fifth tier the pitch height is taken from the corresponding ESPS/waves+ get_f0 file and the following measurements for the pitch range and pitch span analysis according to Patterson (Patterson, 2000) are calculated: mean initial highs, mean subsequent highs, mean lows, mean final lows.

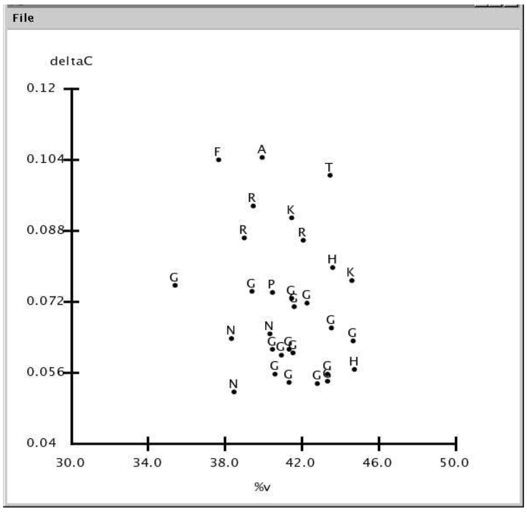

Figure 4: The graph shows an example for the distribution of speech rhythm of British English native speakers (N) and learners of english: G=German, P=Polish, A=Arabic, H=Hungarian, K=Korean, T=Thai, F=French.

A third area of prosodic analysis of the speech data is tonal alignment in stressed syllables. English and German differ in that respect (Grabe, 1998) an it might be a useful feature for the description of non-native prosody. For the analysis, the time stamps on the tone tier, which provides information about the occurrence of stressed syllables, the time stamps on the pitch tier, which gives pitch maxima, and the time stamps on the rhythm tier, which indicates the vowel boundaries, are combined and the presence of pitch height in relation to the vowel boundary is calculated.

Other features important for the analysis of language learners' prososdy such as speech rate, fluency, and intonation patterns are also supported by TASX analysis tools. All tools will be freely available to the scientific community as open source software.

5. Conclusions

Despite the early stage of the research the TASX-based approach has already proved to be highly efficient and reliable. The time consuming task of phonetically analysing speech data is partially substituted by automatic analysis. In the transformation process from non-XML to XML-annotated data some errors in the human annotations can be detected. Furthermore, due to the highly structured format of the TASX-converted data more complex research questions can be investigated in a systematic way.

The very good availabilty of XML aware software and tools enabled us to develop a powerful linguistic environment in a very short time. Even more important, the TASX-annotated data can be transformed into large number of different formats. The will hopefully lead to the creation of linguistic resources which can be used over a long period of time by different researchers with a wide range of scientific goals.

6. References

J. Archibald. 1998. *Second language phonology*. Amsterdam: Benjamins.

S. Bird and M. Liberman. 1999. A Formal Framework for Linguistic Annotation. Technical Report MS-CIS-99-01, Department of Computer and Information Science, University of Pennsylvania.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

J. Boyle. 1987. Perspectives on stress and intonation in language learning. *System*, 15(2):189–195.

J.E. Garcia, U. B. Gut, and A. Galves. 2002. Vocale - a semi-automatic annotation tool for prosodic research. In B. Bel and I. Marlien, editors, *Proceedings of the Speech Prosody 2002 conference, 11-13 April 2002. Aix-en-Provence: Laboratoire Parole et Langage*, pages 327 – 330.

D. Gibbon and T. Trippel. 2001. Pax - an annotation based concordancing toolkit. In Peter Buneman Steven Bird and Mark Liberman, editors, *IRCS Workshop on Linguistic Databases, University of Pennsylvania, Philadelphia, USA*.

E. Grabe. 1998. *Comparative intonational phonology: English and German*. MPI series in psycholinguistics.

U. B. Gut and J.-T. Milde. 2002. The prosody of nigerian english. In B. Bel and I. Marlien, editors, *Proceedings of the Speech Prosody 2002 conference, 11-13 April 2002. Aix-en-Provence: Laboratoire Parole et Langage*, pages 367 – 370.

R. Ihaka and R. Gentleman. 1996. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

J.-T. Milde and U. B. Gut. 2001. An XML-annotated phonetic corpus database. In *Workshop on linguistically interpreted corpora(LINC2001), Leuven*.

D. Patterson. 2000. *A Linguistic Approach to Pitch Range modelling*. Ph.D. thesis, University of Edinburgh.

F. Ramus, M. Nespor, and J. Mehler. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition*, 73:265–292.

T. Schmidt. 2001. Gesprächstranskription auf dem Computer - das System EXMARaLDA. *Gesprächsforschung, http://www.gespraechsforschung-ozs.de*, 2.

K. Silverman, M. Beckman, J. Pitrelli, F. Ostendorf, C. Wightman, J. Pierrehumbert, and J. Hirschberg. 1992. Tobi: a standard for labeling english prosody. In *Second International Conference on Spoken Language Processing 2, Bannf, Canada*, pages 867–870.

W. N. Venables and B. D. Ripley. 1999. *Modern Applied Statistics with S-Plus. Third Edition*. Springer. ISBN 0-387-98825-4.