ADABA - Aussprachedatenbank Deutsch. The Pronounciation Database of the National Standard Varieties of German

Rudolf Muhr
Karl-Franzens University, Graz, Austria
E-mail: muhr@gewi.kfunigraz.ac.at

1. ABSTRACT

This paper gives a comprehensive report on major the results of the ongoing project „Varieties of Austrian German – Standard pronunciation and varieties of standard pronunciation“ which started in January 2001 and will be finished in 2004. It is financed by the research fund of the Austrian national bank and supported by the national Austrian broadcasting corporation ORF. One of the major results is the ADABA database which contains the pronunciation of a corpus of 13.500 phonetically rich words spoken by six model speakers coming from Austria, Germany and Switzerland totalling 81.000 single words. The project, the database and the ADABA-Transcriber which has also been developed in the course of the project will be presented in detail and an outline of its possible use for phonetic research will be given. The acronym ADABA stands for: AUSSPRACHE-DATENBANK-AUSTRIA (Pronunciation-Database-Austria.)

2. The Background of ADABA-Database

The ADABA-Database has been developed in the context of the above mentioned project „Varieties of Austrian German – Standard pronunciation and varieties of standard pronunciation“ whose main objective is to develop an Austrian German Pronouncing Dictionary (AGPD) on a theoretically and empirically firm basis. The AGPD is a codification attempt of standard forms of the pronunciation of Austria German (AG) providing reliable information about the so called „standard pronunciation“ (and other common pronunciations) prevailing in Austria and contrasting it to the standard pronunciation of the two other German speaking countries. The main focus is however on the description of the so called "target-norm" or "model pronunciation". The differentiation of other levels of pronunciation which are eventually common throughout the country but not acknowledged as model pronunciation is sought. The dictionary therefore intends to provide some information on non-model forms of pronunciation and their usage, making users of the dictionary understand that in modern societies different levels of the so called „standard language“ are co-existing. These non-codified forms of pronunciation often function as a „standard“ in certain domains, regions or speech situations. The realisation of these objectives however depends on the further funding of the project.

3. The AGPD-Project - Objectives and participating institutions

3.1 The participating institutions: The AGPD-project is carried out by the “Austrian German Project” at the Department of German of Graz University. Partners in the project are the “Institute of Electronic Music and Acoustics” at Graz University for Music and Arts and the ORF - Austrian National Broadcasting Company. The project started in January 2001 and will be finished by September 2004.

3.2 The objectives of the project[1]:

1) Building a representative pronouncing dictionary of AG comprising information on AG pronunciation contrastive to the other NAV.
2) Building a phonetic database (ADABA) which contains: A) Reading pronunciation of a word list containing 13.500 words read by 1 male and 1 female model speaker from Austria, Germany and Switzerland resulting in 81.000 sound files. All files have been transcribed and the transcription of all words and texts is shown by the interface of the database.
B) Reading pronunciation of 2 texts: a) A literary text (1014 words) and b) a news text (914 words) read by the 6 model speakers of the 3 NAVG. These data are intended to provide information on important features of fluent speech (process phonology) produced by model speakers.
C) The sound data of a list of 393 phonetically rich words read by 54 additional model speakers from Austria, 10 from Switzerland and 22 from Germany. This word list was read by 3 male and 3 female professional radio announcers coming from each regional station of the Austrian Broadcasting Corporation in the 9 Austrian "Bundesländer" (regions). The data of these speakers is intended to provide more empirical data and to deepen the knowledge about the pronunciation of model speakers coming from different areas of Austria and the other German speaking countries.
D) About 100 sound files, 2-3 minutes in length, containing fluent speech of both model speakers and untrained speakers speaking on Austrian radio or television. These sound files are intended to provide information on features of fluent speech.
The phonetic database in its final stage will contain about 110.000 sound files and give a thorough documentation of

[1] Until now (spring 2003) the tasks 2A), 2D) and 3) have beeen completed. The data for 2B) and C) have been collected. Task 4) and 5) are still to be done.

ISBN 1-876346-48-5

different levels of pronunciation in Austria (and the other German speaking countries).
3) Providing a user-friendly interface for searching the database, listening to the sound-files and analysing the sound files with a tool for acoustic analysis. (This interface is presented below.)
4) Defining reliable rules for the AG-pronunciation and applying them to a lexicon of 50.000 words which then forms the content of the AGPD.
5) Publication of the AGPD in printed version together with the phonetic database on DVD. The database will partly also be made available on Internet.

4. The theoretical framework of the AGPD and the ADABA-Database

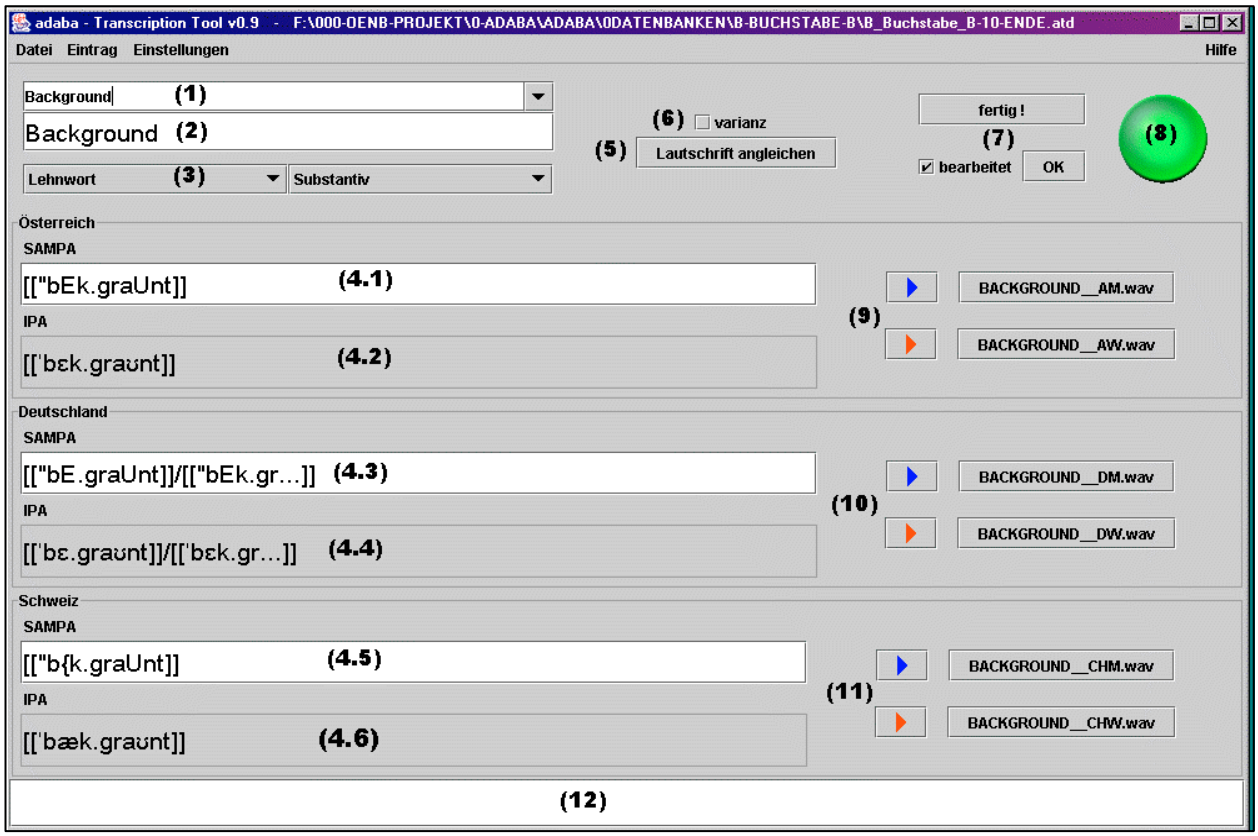
4.1. German as pluricentric language - Austrian German as a National Variety of German

The starting-point of the AGPD-project is the sociolinguistic theory of pluricentric languages as developed by Kloss (1953/1978), Clyne (1992) and others. Pluricentric languages are marked by their spread over several countries and their partial differences in linguistic and pragmatic norms. They are neither a language of its own nor a dialect as the norms of a national variety (NAV) form part of the social and linguistic identity of the populations of the nations sharing a common language. The differences in norms usually are spread over all levels of the linguistic structure, although the ones in pronunciation and lexicon usually are most numerous and those speakers are most aware of. Clyne (1992) pointed out that in pluricentric languages "dominating" and "other" varieties have to be distinguished. The codification of norms usually is only taking place in linguistically dominating countries. This leads to dictionaries whose norms are partly foreign to the usage in the "other" nations which contributes largely to the wide-spread impression that the pronunciation forms of non-dominating countries are not "standard". The AGPD therefore attempts to provide a codification of the Austrian standard norm of pronunciation contrastive to the norms in the two other German speaking countries. A cornerstone to achieve this objective is the distinction between "national" and "regional" norms which exist in Austria. Considerable effort had to be spent to find speakers whose pronunciations are highly accepted across all layers of society and in all regions of the country.

4.2. The methodological steps for the creation of the AGPD

Step (1): The development of a theoretical framework for the definition and choice of pronunciation variants: A new theoretical framework for the choice of linguistic varieties was developed. It assumes (a) that the "media presentation norm" is central for the definition of linguistic "target" norms in modern post-industrial societies and (b) the codification of the "standard" pronunciation should be based on this norm. Finally, (c) three major pronunciation domains were defined:
(A) Pronunciation domain (1) is defined as (a) reading pronunciation which is (b) prepared, (c) monologue or read dialogue, (d) based on written language, (e) public, (f) fact-directed and (g) socially distant.
(B) Pronunciation domain (2) is defined as (a) the pronunciation of free speech which can be (b) prepared and produced in freely spoken lectures, reports or presentations on TV or radio or (c) unprepared in discussions in front of a TV-public or on radio, (d) is always dialogue, (e) only partly based on written language, (f) public, (g) fact and person-directed and (h) socially distant or semi-distant;
(C) Pronunciation domain (3) is the pronunciation of (a) free speech which is (b) always dialogue, (c) not based on written language, (d) semi-public or private, (e) person-directed, (f) socially close. In any of the three domains trained and untrained speakers can be observed and their pronunciations analysed.
The codification of the AG model-pronunciation is primarily based on the reading pronunciation of selected trained speakers working as professional announcers on the Austrian national Broadcasting Corporation ORF. They are realising language in pronunciation domain (1). Data of pronunciation domain (2) are to be included in the dictionary. The amount of this data is however dependent on the further funding of the project.
Step (2) - The selection of the Austrian model speakers: The selection of the Austrian model speakers was done in two steps: In a first round 35 announcers of the ORF were asked to read a corpus of 392 phonetically rich words. Their realisations were judged by linguists and phoneticians and 8 male and 9 female speakers chosen. In the second step the 17 speakers were subjected to a hearer test via a web-questionnaire with 492 people taking part in the test. The two speakers with the statistically most significant values were chosen for the recording of the model-lexicon.
Step (3) - The selection of the German and Swiss model speakers: The speakers of the two other German speaking countries were chosen by the recommendations of the chief announcers of the Swiss radio corporation DRS and the German station SWR who were both supporting the project.
Step (3): The preparation of the model-lexicon and the model-texts: The model lexicon was derived from four sources: (1) A list of 5.500 communicatively relevant words. This basic vocabulary list was developed by the author of this paper for the "Austrian Language Diploma - ÖSD" forming the basis for the testing of learners of German as a foreign language. (2) A list of the 10.000 most frequent words provided by the “Projekt Deutscher Wortschatz” [Project German Lexicon] situated at the University of Leipzig. It is based on a corpus of 200 mio. running words. (3) The list of 1540 phonetically rich lexical items used in W. Königs (1989) "Atlas zur Aussprache des Schriftdeutschen in der Bundesrepublik Deutschland". [Atlas of the pronunciation of written German in the Federal Republic of Germany]. (4) A list of 3700 highly frequent loan words. All double entries were removed, leading to a list of 13.500 words. For the model-texts a four pages long literary text by Ingeborg Bachmann "The Sphynx" and a text of daily news published on the videotext pages of the ORF was chosen.

ISBN 1-876346-48-5



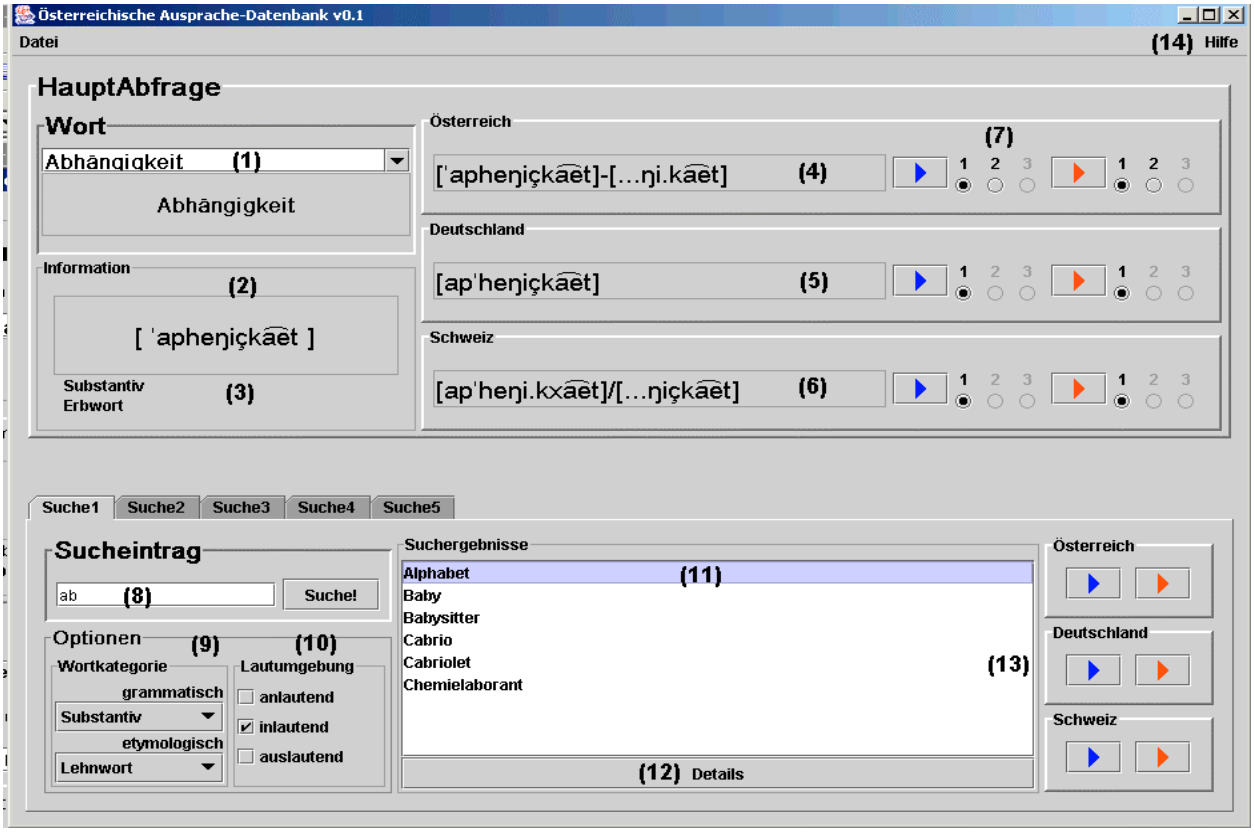
Step (4): The recording of the model-lexicon and the model-texts: The model-lexicon and the model texts were recorded on DAT, processed and later saved on CDROM as 44 kHz, 16 bit mono files. The large sound files were later cut turning each word into a into single sound file.

Step (5): The transcription and the development of a new transcription tool: The corpus of 81.000 words and the model texts have been transcribed in IPA. For the transcription a new tool had to be developed as all known phonetic programmes could not meet the demands. The result is the **ADABA-Transcriber** which allows the typing of the transcription in SAMPA which is then automatically transformed into IPA, saving the data in a database. Every word can be annotated with grammatical and etymological information and the associated sound files listened to as often as necessary. A screenshot of the tool is presented the figure above.

In addition to that, a list of "pretranscribed" words in orthographic and SAMPA-format can be imported into the tool. Typing in the first letter in field (1) will show all words with this letter in initial position. The orthografic form can be edited in field (2) and the new word saved. Linguistic and etymological information can be added with the field (3). The transcription can be entered in SAMPA in the fields 4.1, 4.3 and 4.5., additional information in field (12). The IPA form will be shown in the fields below. The sound files can be listened to by clicking on the buttons (9)-(11). The transcription in field 4.1 can be transferred to the other fields by clicking on button (5). The transcription is finished with button (7), saving everything into the data base and turning button (8) from red into green. Klicking on button (6) will mark the entries with the feature "variance" and showing that several pronunciations of the same word exist. The trancriber also has a help function (14) showing the SAMPA-IPA table. The ADABA-Transcriber will be made available to the scientific community after the final version has been developed.

Step (6): The development of the ADABA-Database - The user interface (see next page for a screenshot): Finally the transcribed data were loaded into the ADABA-Pronunciation-Database which is identical to the one of the ADABA-Transcriber. For the presentation and quering of the data, a special user interface was developed. A screenshot of the interface is shown on the next page. The ADABA-Database is now available in a beta-version allowing searches of different kinds which are presented in detail further down. The interface and the database itself will be enlarged, providing features for the presentation of the data of the additional 86 speakers. There will also be a feature which allows the aligned presentation of the model-texts. At present two kind of searches can be conducted in the ADABA-Database:
(1) A main search by typing in words (or parts of words) in field (1). After hitting the enter-key all information connected with the lexicon entry will be shown in field (3) presenting the pronunciation of the Austrian speakers. (In this version only the data of the male Austrian speaker will appear as the rest still has to be implemented.) Parallel to

ISBN 1-876346-48-5 © 2003 UAB



these data, the transcription of all 6 speakers is shown in the fields (4)-(6). The sound files which are presented automatically, can be listened to via the buttons on the right hand side (7). If there are different pronunciation-variants they can be listened to by clicking on the numbers to the right of the play-buttons.

(2) Specific searches can be conducted by typing words into field (8). The search can be specified in respect to word classes and etymological categories (native word, loan word, proper noun) and the phonetic context (initial, medial, final). The words corresponding to the chosen features will appear in field (11). They can be listened to by marking a word and by clicking on one of the play buttons on the right (13). A click on field (13) will show the transcription of all speaker realisations in field (4)-(6) and also allow the playing of the sounds of each speaker.

5. Outlook

The project is now in its final phase. The model texts and the lexical data of the additional 84 speakers have to be included into the database and the GUI adapted. After the phonetic analysis of the data, the rule set will be applied to a lexicon of about 50.000 words. The AGPD will then be generated and in a last step checked by actual pronunciation data in the media. During the transcription process about 40 pronunciation rules of AG already emerged from the data. They will have to be verified by a thorough acoustic analysis which is ongoing at present. Already in course of the transcription-process of the single words-corpus many theoretical and practical problems of assigning the correct IPA-letters to the sounds emerged. The main problem was that there are often little differences in the phonetic substance of the pronunciation but enough to be percieved by speakers and identified as foreign to them. A printed version and the ADBA-Database on DVD will be published in 2004. We hope to provide a tool which is useful for phonetic research as well for the training of professional speakers and in language pedagogy.

REFERENCES

[1]. Clyne, M. 1984. *Language and Society in the German-Speaking Countries*. Cambridge. CUP.

[2]. Hollmach, Uwe (1996): *Soziophonetische Grundlagen zur Neukodifizierung des Aussprachewörterbuches*. In: Hallesche Schriften zur Sprechwissenschaft u. Phonetik", Band 1. Verlag Werner Dausien Hanau und Halle. ISBN 3-7684-6538-11996.

[3]. Kloss, H. 1978. *Die Entwicklung neuer germanischer Kultursprachen seit 1800*. 2. enl. ed.. Düsseldorf: Schwann.

[4]. König, Werner (1989): *Atlas zur Aussprache des Schriftdeutschen in der Bundesrepublik Deutschland*, 2 Vol. Ismaning: Hueber.

[5]. Stubkjaer, F. T., 1995. Überlegungen zur Standardaussprache in Österreich. In Muhr, R. et. al. (eds.), 1995. *Österreichisches Deutsch*. Wien. öb&vhpt. S.248-268.

[6]. Takahashi, H., 1996. *Die richtige Aussprache des Deutschen in Deutschland, Österreich und der Schweiz nach Maßgabe der kodifizierten Normen*. Peter Lang Verlag, Frankfurt/M..