GENIE: The Corpus for Spoken Lower Sorbian (GEsprochenes NIEdersorbisch)

Article · January 2010

CITATIONS
2

READS
172

3 authors, including:

Roland Marti
Universität des Saarlandes
34 PUBLICATIONS **33** CITATIONS
SEE PROFILE

Bistra Andreeva
Universität des Saarlandes
87 PUBLICATIONS **552** CITATIONS
SEE PROFILE

GENIE: The Corpus for Spoken Lower Sorbian (GEsprochenes NIEdersorbisch)

Roland Marti, Bistra Andreeva, William J. Barry
Department of Slavonic Languages, Saarland University, Saarbrücken, Germany
Phonetics, Saarland University, Saarbrücken, Germany
e-mail: rwmslav@mx.uni-saarland.de, andreeva@coli.uni-saarland.de, wbarry@coli.uni-saarland.de

Abstract
Lower Sorbian is a Slavonic minority language spoken in Eastern Germany in German-speaking surroundings. The language is on the brink of extinction as there are basically no native speakers below the age of sixty. Therefore, the documentation of spoken Lower Sorbian is crucial. The corpus of spoken Lower Sorbian GENIE (GE[sprochenes] NIE[dersorbisch]: http://genie.coli.uni-saarland.de/) is the first documentation of this kind. It brings together various kinds of spoken Lower Sorbian: recordings from the archive of Sorbian broadcasts (years 1956-2006), recordings from the Archive of Sorbian Culture (dialect recordings 1951-1971), and new recordings from native speakers made especially for the corpus in 2005/2006.

The paper presents the corpus and its defining features, paying special attention to the particular situation of Lower Sorbian and its bilingual speakers. On the one hand, there is a very strong German influence; but on the other, Upper Sorbian interference is also clearly recognizable in the recordings. Furthermore, the paper illustrates the problem of what constitutes the speech of a native speaker in the case of minority languages. Finally, the problems of corpora of endangered languages are discussed.

1. Sorbian

Sorbian is currently geographically the furthermost western part of the Slavic speaking area. It is at present a language island (more exactly, an archipelago of islets) within a German speaking area, that is situated in Upper and Lower Lusatia. This represents the remainder of the originally much larger territory, which, by means of language exchange, was gradually Germanized; a process that was repeatedly triggered and fostered by language-political measures that still continue (cf. Figure 1).

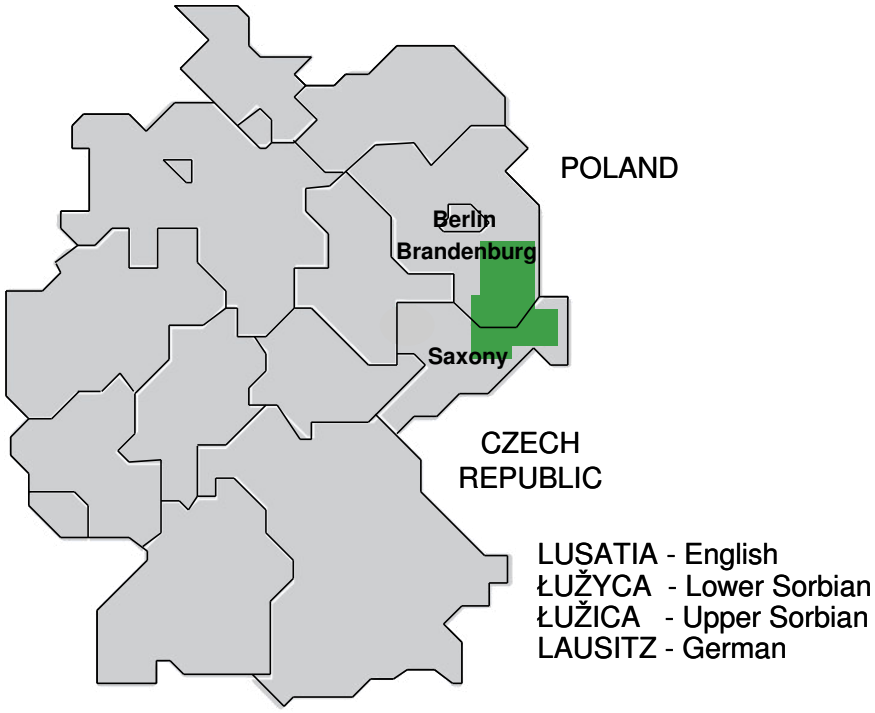

Figure 1: The Sorbian-speaking region in Germany.

This language area can be roughly divided into Upper and Lower Sorbian. Only in the Upper Sorbian area, more precisely in the Catholic districts, are there still villages where Sorbian is the common language (Scholze, 2008); elsewhere it remains nothing more than a family language, or rather the language of the older generation(s). The number of people with an active command of Sorbian can only be estimated. The estimates vary between 15,000 and 30,000 for Upper Sorbian and between 5,000 and 10,000 for Lower Sorbian (Jodlbauer, Spieß & Steenwijk, 2001). Upper, as well as Lower, Sorbian are autonomous languages. They are officially acknowledged as minority languages in Germany, first, in the constitutions and appropriate laws concerning Sorbs (or Sorbs/Wends) in the Free State of Saxony and the state Brandenburg[1] and, second, in the European Charter for Regional or Minority Languages.

The main problem for the Sorbian language is the dying-off of the Sorbian speaking community due to the lack of younger native speakers and the consequent shrinking of the area in which Sorbian is spoken. Geographical shrinkage is a phenomenon that has been observed since the 16th century. Both trends have been accelerating since the mid 19th century, and neither the revival measures nor fostering throughout the German Democratic Republic era could stop them. There are language preservation and revitalization measures at present

[1] The official name in Brandenburg is "Sorbs/Wends" ("Sorben/Wenden") and "Sorbian/Wendish" ("sorbisch/ wendisch") since a part of the Lower Sorbian speaking community refuses the name "Sorbs" ("Sorben") and "Sorbian" ("sorbisch"), where native speakers are concerned. According to linguistic (Slavic) tradition only "Sorbs" ("Sorben") and "Sorbian" ("sorbisch") are used.

(especially the so called WITAJ-project; Budar & Norberg, 2006), which can, however, at best slow down the language assimilation process. The situation of Lower Sorbian is particularly dramatic since inter-generational transmission does not exist any longer and children are led by means of (partial) immersion to the status of a kind of "secondary native speaker".

There are yet other specific problems concerning Lower Sorbian. The revival of Sorbian life and its organization after the Second World War was primarily initiated in the Upper Sorbian region and by Upper Sorbian exponents. This led to the perception that the cultural life was Upper-Sorbian oriented, which was in fact partially the case. This was experienced especially intensively in the language domain. The spelling reform from 1949-1952 brought about the approximation of Lower Sorbian to Upper Sorbian orthography. Since pronunciation that oriented itself on the written language was fostered and required at school and in the media, the spelling reform also had orthoepic consequences (so-called "spelling pronunciation"). The Upper Sorbian linguistic influence was further strengthened by the fact that, owing to the small number of autochthonous Lower Sorbian experts, functionaries in Sorbian organizations and teachers came predominantly from Upper Lusatia, and their language did not conform to the linguistic features of Lower Sorbian. This resulted in the popular impression that the Lower Sorbian standard language does not represent real Lower Sorbian at all, but an overall Sorbian hybrid language at best, or a kind of Upper Sorbian that had been adjusted slightly to Lower Sorbian. Many native speakers of Lower Sorbian therefore refused to participate in official efforts to strengthen the language and restricted its use to private life. Often they even stopped transmitting the language to the next generation. On the other hand, the official language policy, centred on the standard language and neglecting dialects, gave rise to the feeling in Lower Sorbian speakers that they could not speak correct Sorbian (an opinion that is heard repeatedly during field recordings). This explains the wish for reinforced demarcation from Upper Sorbian which emerged when state control over cultural life ceased. The latter finds expression in the adoption of different terminology ("Wendish" instead of "Lower Sorbian", cf. n. 1), in the withdrawal of some parts of the spelling reform from 1949-1952, and in the rejection of a purist language that is felt to be Upper Sorbian.[2]

2. The Corpus for Spoken Lower Sorbian GENIE

In view of the precarious situation of Lower Sorbian that was described in relevant studies (Jodlbauer, Spieß & Steenwijk, 2001; Norberg, 1996), it was foreseeable that the "authentic" mother tongue would no longer exist within one generation at best. That turned out to be particularly fatal for the spoken language since the

[2] This results in the current (re)appearance of lexical Germanisms (*lazować* instead of *cytać*, *hundert* instead of *sto*), that have always been in colloquial use, also in written language. The similar situation can be observed in the grammar section, e.g. with determination (occasional use of the definite and marginally also the indefinite article).

"secondary mother tongue" (the maximum goal aimed at by efforts of revitalization) differs strongly from the "authentic" mother tongue, especially in its pronunciation.[3] In this respect, it was important, and extremely urgent, to document spoken Lower Sorbian. With this objective in mind, the corpus GENIE: GEsprochenes NIEdersorbisch (Spoken Lower Sorbian) was created. The corpus creation was partially funded by the Scientific Committee of the University of Saarland in the years 2005-2006. The endeavour was also financially supported by the Radio Berlin-Brandenburg (RBB) and the Sorbian Institute/Serbski Institut. In order to make this corpus internationally usable for the scientific research, it was made available on the web (http://genie.coli.uni-saarland.de). The GENIE website is supported by the Insitut für Phonetik (http://www.coli.uni-saarland.de/groups/WB/Phonetics/index.php) together with the Institut für Slavistik (http://www.uni-saarland.de/fak4/fr44/) at the University of Saarland. Due to copyright and data privacy protection rights, it could not be made generally available; its use is permitted for scientific purpose by application (http://genie.coli.uni-saarland.de/cgi-bin/benutzer.html). The corpus arrangement was structured to meet the special features of the situation of Lower Sorbian presented above and, where possible, to take into account the diachronic level.[4] There are more than sixty hours of spoken Lower Sorbian in its distinct variants available in GENIE. Even though the period of time covered by the recordings ranges only from 1951 to 2006, the speakers' dates of birth indicate that the diachrony is considerably deeper: the oldest speaker was born in 1860 (he was 94 years old at the time of the recording), the youngest speaker was born in 1973. Individual diachrony is also traceable since several people are represented in multiple recordings that were produced at different times.

2.1 Sources

The corpus consists of recordings from three different sources:

a) Archive of the Sorbian Radio (Studio Cottbus of the Radio Berlin-Brandenburg RBB, formerly ORB, earlier still Radio of GDR)

This source consists of 110 recordings made between 1956 and 2006. Speakers of dialects and of the standard language (native speakers of Lower Sorbian/ Wendish, Upper Sorbian or German) are both represented in different variants of the standard language. The text types are very different: conversation, interview, address, report etc.

[3] The reason for this is primarily due to the fact that the teachers employed in the revitalization project WITAJ, apart from a few exceptions, do not have a command of Lower Sorbian as their mother tongue, but at best as their secondary mother tongue.

[4] Owing to copyright, the oldest recordings of Sorbian could not be adopted from the Berlin Archive, therefore only marginal diachronic depth is taken into consideration: the recordings were made in the years 1951-2006.

b) Archive of Sorbian Culture/Serbski kulturny archiw (SKA) in the Sorbian Institute/Serbski Institut
The source contains 135 recordings made between 1951 and 1971.The recordings were compiled for linguistic purposes by the Institute, in particular for the Sorbian Linguistic Atlas (cf. References SSA 1-13 1965-1993). Its aim was the recording of local dialects (story, interview, elicitation etc.).

c) The field study project specifically for this corpus
The source consists of 100 recordings made between 2005 and 2006. They involve conversations between J. Frahnow (pastor and native speaker) and mostly elderly native speakers whose speech usually represents a local dialect. While selecting the recordings and test persons, we attempted to depict the complexity of dialectal forms of Lower Sorbian/Wendish along with diverse standard linguistic variants employing the three sources mentioned.

2.2 Metadata files

There is a data record sheet for every recording containing the most important information about the recording. Specifically, these are:

- call number (the recording identifier): this consists of the letters f, r or s and a four-character-number where f means field recording created by J. Frahnow, r stands for recordings from the radio archive of the RBB, and s signifies recordings from the Archive of Sorbian Culture. In addition to the call numbers valid for this corpus, there are archive call numbers as used in the source.
- text type (e.g., conversation, interview, report)
- contents (e.g., village life, customs, farming)
- place of the recording
- date of the recording
- indication of sex (names are not given to protect the person's identity)
- speaker's place of birth
- speaker's date of birth
- dialect
- family language: it is specified here whether the family language was Lower Sorbian/ Wendish, German or mixed (or Upper Sorbian where applicable)
- places of residence
- education

The place names in the arrays (place of the recording, place of birth, dialect and place of residence) are given in German and Lower Sorbian/Wendish and can be shown and arranged in three sections: place, municipality, and district. Additionally, all the Lower Sorbian places covered are allocated to the dialect

areas. In doing so, the classification of the Sorbian Language Atlas was taken into consideration, which ultimately goes back to the categorization by Muka (1911-1926). In it, only Lower Sorbian dialects proper or transitional dialects are distinguished. In the case of native speakers of Upper Sorbian, there is only a reference to this fact without indication of the dialect area. In case of non-native speakers or native speakers that use the standard language, the word "standard" is used.

There are several metadata sets available for some recordings, namely in cases where there is more than one speaker participating in the recording (hosts and interviewers were usually not taken into account). The call numbers of the metadata sets are identified in these cases by the attached index letters (e.g., a, b, etc.).

Access to the datasets and audio recordings in the corpus may be obtained either directly, by stating the call number, or indirectly by using a search form, within which you can search or classify all specified arrays with intelligent filter functions.

2.3 Technical data of the recordings

In addition to the specified background information, data record sheets comprise the following information:

- length of the recording in minutes and seconds
- size of the .wav-file in bytes/kilobytes/megabytes
- size of the .mp3-file[5] in bytes/kilobytes/megabytes
- sampling rate in Hz
- amplitude quantization rate in bits per sample
- number of channels (1 for mono, 2 for stereo)
- signal-to-noise ratio SNR (as yet only with data from the field search project)
- bit rate (.mp3-file) in kBit/s

3. Examples from GENIE

It is evident from the description of the GENIE corpus that the material can be analysed with various objectives in mind. For one thing, the description and the comparison of the structural characteristics of the various dialect areas are an attractive challenge in itself. Even though the spontaneous speech of the recordings does not allow for an exhaustive grammatical description, the newly recorded material provides a valuable supplement to the (not immediately accessible) dialect recordings made during the German Democratic Republic era. Another important question is to what extent the spoken standard language may vary and, depending on the speaker's origin, adopts a dialectal form, thus actually containing Lower Sorbian, Upper Sorbian or German features. The focus of our

[5] mp3 audio files are highly compressed in size. They take much less time to transmit over the internet.

first analyses, though, will be on the influence of German on spoken Lower Sorbian; an influence that grew steadily over the 20th century, but which had been present a long time before. The comparison of recordings of younger and older people can shed light on the extent of this influence, as well as on the linguistic features affected by it. More striking yet is the comparison of recordings of the same person made at different times.

According to the existing descriptions (Schwela, 1906; Janaš, 1984; Starosta, 1991), there are well-known phonetic dissimilarities between German and Lower Sorbian on the segmental level, the vowel quality and quantity, the R sound, the realization of plosives with regard to voicing and aspiration, as well as the existence of the dark L or a [w] and of the correlation of palatalization, widespread in Slavic languages. There are, above all, characteristic features of intonation and word stress known from impressionistic descriptions of the prosody. Other rarely mentioned, though important discrepancies, are word-chaining modes, such as the division of neighbouring vowels by means of a glottal stop or the type of voice assimilation (progressive or regressive).

As examples of the existing and growing impact of the influence of German on Lower Sorbian, we show here four of the phenomena mentioned above in utterances of an elderly speaker (A, born in 1890) and of a younger speaker (B, born in 1960).

Figure 2, a representation of the microphone signal and the spectrogram of the utterance, "Chtož tu rolu wobźěłajo" (English "Who works on the land") illustrates several pronunciation features in one short stretch of speech that prove the influence of German, three of which we comment on below:

1. In the word "rolu" the /r/ is realized as a uvular approximant [ʁ′](see I).
2. "wobźěłajo" /'obʑewajo/ starts with a glottal onset instead of a smooth transition from "rolu" (see II) or an alternatively possible [h]
3. The syllable-final /b/ and the following syllable-initial /ʑ/ are voiceless (see III).

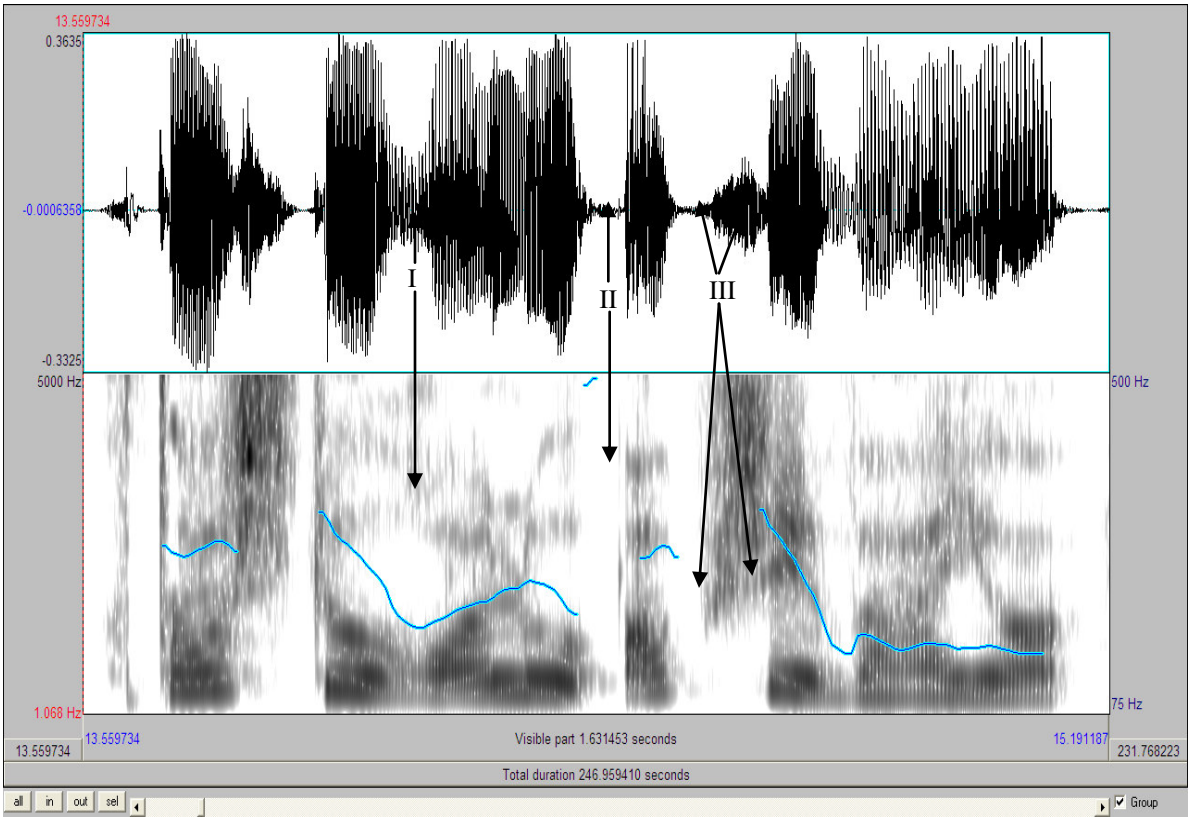

Figure 2. The utterance "Chtož tu rolu wobźěłajo" (here: [xtɔʃ tʊ ʁɔlu ʔɔpʃevajɔ]) by speaker B (born in 1960) with (I) uvular [ʁ], (II) hard vowel onset (glottal stop) und (III) devoicing at the word coda with progressive devoicing of a voiced initial fricative.

In Figure 3, depicting the oscillogram of an acoustic time signal and the spectrogram of the utterance "tak daloko" (English "so far"), the voiceless plosives /t/ (see I) and /k/ (see II) demonstrate, contrary to the claim that in Lower Sorbian voiceless plosives are unaspirated, clear features of a moderate degree of aspiration (in both cases 26 ms). The measured duration of aspiration is relatively short if compared to that of monolingual speakers of German. Therefore it is important to examine whether an intermediate form (similar to the weak aspiration with Canadian speakers of French; Sundara et al., 2006; Fowler et al., 2008) has become established in Sorbian, within this generation or with this speaker alone.

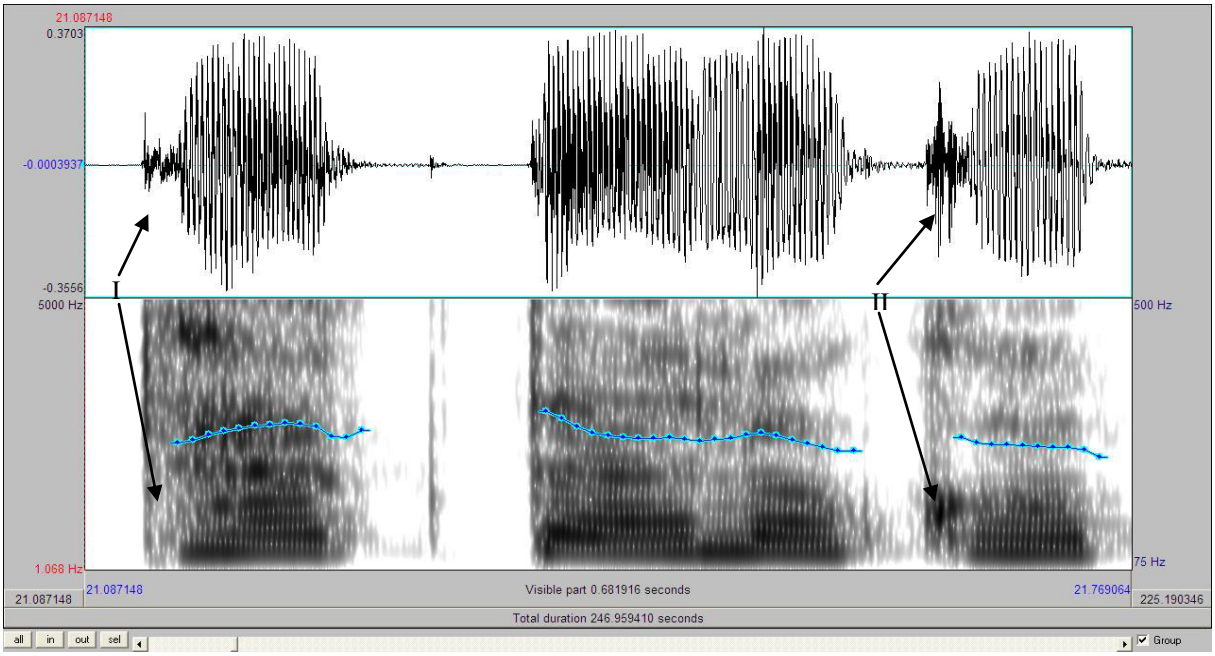

Figure 3. The utterance “tak daloko” (here: [t^{h}ak dalɔk^{h}ɔ]) by speaker B (born in 1960), where clear aspiration (I) of /t/ and (II) of /k/ can be noticed.

The older speaker (born in 1890) demonstrates a different articulation pattern. Indeed, in figure 4 in her statement, “To njejo tak dobre” (English “It’s not that good”), a tendency to aspirate can be observed: /t/ in “to” manifests an aspiration duration of 37 ms (see I).

On the other hand, following /k/ in “tak” she produces a fully voiced initial /d/ in “dobre” that affects /k/ regressively, making it voiced (see II). This suggests that the assimilation process contrasts with the common German pattern but corresponds to what is typical of other Slavic languages. The apical [r] in “dobre” also differs from the German standard-/r/, which is a uvular fricative [ʁ]. There are two signal muting taps of apical [r] to be seen in spectrogram as well as in the microphone signal (see III).

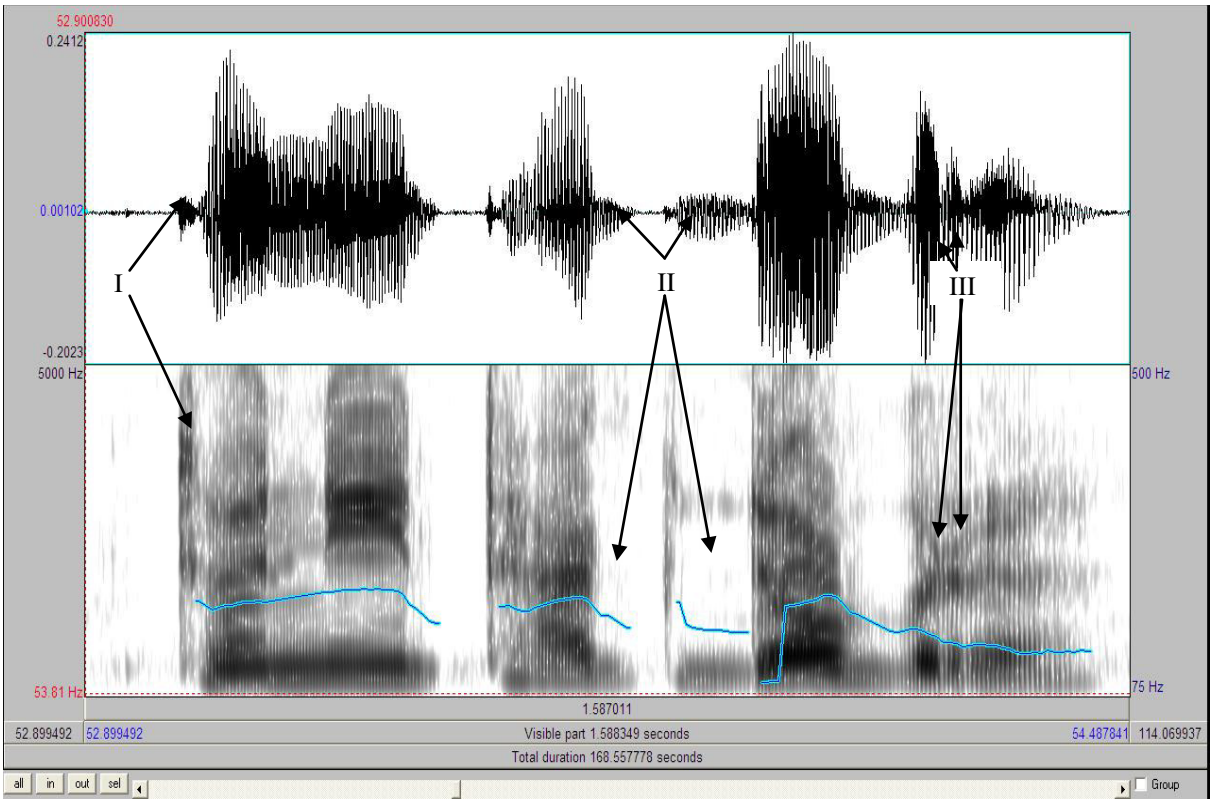

Figure 4. The utterance "To njejo tak dobre" (here: [t^{h}ɔ ne t^{h}ag dɔbrə]) by speaker A (born in 1890), where (I) aspiration of /t/, (II) a fully voiced /d/ with partial voicing of the preceding /k/ and (III) a double-contact apical /r/ can be observed.

As far as the fourth phenomenon in the younger speaker's recording is concerne (the missing smooth transition from one vowel to the next across a word boundary), it cannot be maintained that in earlier times glottal constriction, according to German pattern, did not appear. In a short utterance ("a to ak," English "as") of speaker A, there is a clear glottalization at the beginning of the utterance and at the word boundary between "to" and "ak" (see I and II in figure 5). Further studies will allow us to determine how often such instances of glottalization occur in her speech. It also cannot be ruled out that Slavic languages behave similarly to other "binding" languages (French, Italian, English etc.) and dialects (such as Alemannic). That is to say a stressed word with an initial vowel in an emphatic context can very well start with a hard glottal onset. In the younger speaker's example, however, the glottalization appears in non-emphatic context. The older speaker's utterances are characterized by a general emphatic "word by word" style. The utterance is not distinctively emphatic, but the glottalization might be attributed to this general style. A further uncertainty, when comparing the two speakers, results from age-related differences in the voice quality that add to the difficulty of interpreting glottal phenomena.

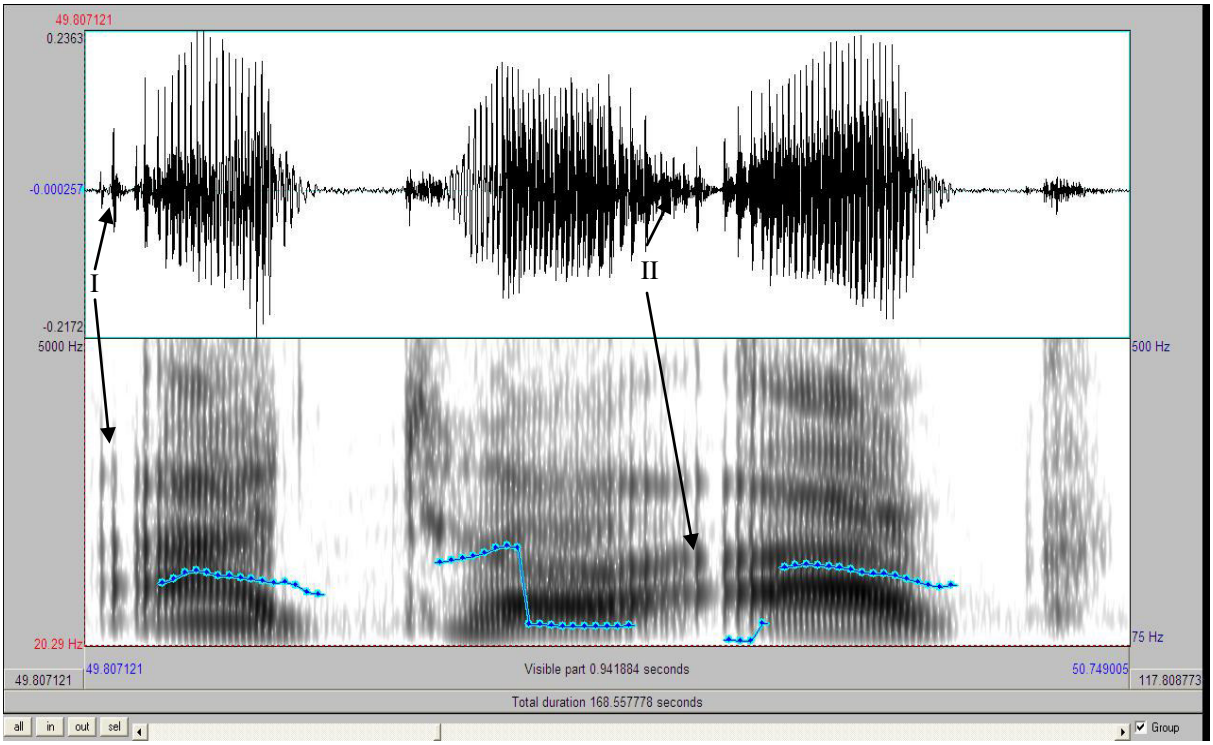

Figure 5. The utterance "a to ak" (here: [ʔa tʰɔ ʔak]) by speaker A (born in 1890) with glottalization (I) at the beginning of the utterance and (II) at the word boundary between "to" and "ak."

4. Corpora of endangered languages – an exceptional case?

Following the presentation of a concrete corpus of an endangered language, we should ask whether, from a general linguistic perspective, corpora of endangered languages, or of micro languages in the broader sense (see The UCLA Phonetics Lab Archive [http://archive.phonetics.ucla.edu/], The Endangered Language Fund [ELF: http://endangeredlanguagefund.org/], DOkumentation BEdrohter Sprachen/documentation of endangered languages [DOBES: http://www.mpi.nl/DOBES/], and the Leipzig Endangered Languages Archive [LELA: http://www.eva.mpg.de/lingua/resources/ lela.php] among others), are essentially different from the corpora of other languages and whether this has consequences for their planning, composition and supervision. In fact, there are differences, but they are not of a principal nature.

An important difference concerns information value or, in other words, representativeness of the corpora. Paradoxically, the corpora of endangered languages are simultaneously more and less representative than those of other languages. The higher degree of representativeness becomes especially clear in the case of written corpora. Only languages with a limited written tradition may include a high percentage of all that has been written in the corpus.

There are two reasons for lower representativeness. First, endangered languages are either not documented at all, or if they are, then by relatively small-sized corpora and only rarely by means of several corpora. In addition, the data that exists has usually been collected by chance and does not reflect an intentional selection. The second reason for lower representativeness lies in the fact that the norm of endangered languages is less fixed, and so there is greater variability within them that can only be imperfectly represented. It is even possible that idiolectal predominance in a corpus may distort linguistic structures.

A further discrepancy is related to the composition, processing and supervision of the corpora. As far as endangered languages are concerned, the group of people that are interested in the corpora and are capable to put them together is rather small. The same applies to the financial possibilities of minorities. As a consequence, corpora of minority languages, if they are created at all, cannot be specialized (they are the proverbial 'all-in-one' tools) and will only be partially annotated, if at all. Continuous development, updating and documentation are only possible to a very limited degree.

A major difference is ultimately inherent in the function of the corpora. As far as endangered languages are concerned, the corpus is not a linguistic working tool in the first place. It is, rather, a memorial with a quite distinct culture-political objective. It shall document what still exists and what will possibly soon disappear.[6] This may well have consequences for the choice of the texts to be recorded if the "antiquarian" idea prevails.

Corpora of endangered languages are clearly an exceptional case. Both producers and consumers must take this into consideration. The producers must take into account the limiting general conditions and the additional functions and ensure that such corpora will be supervised in spite of limited resources. The users must show understanding for the particularities of such corpora and also be willing to contribute actively to their optimization, for example, by making the transcriptions and annotations they created themselves available for the corpus.

References

Budar, L. & Norberg, M. (2006). „Les écoles sorabes après 1990". Education et Sociétés Plurilingues 20 (juin): 27-38.

Fowler, C. A., Sramko, V., Ostry, D. J., Rowland, S. & Halle, P. (2008). Cross-language phonetic influences on the speech of French-English bilinguals. *Journal of Phonetics* 36, pp. 649-663.

Janaš, Pětr (1984). *Niedersorbische Grammatik für den Schulgebrauch*. Bautzen: Domowina.

[6] It is not a coincidence that in the "Archive of vanished places" ("Archiv verschwundener Orte/archiw zgubjonych jsow") in the village of Baršć/Forst, recordings of Sorbian language are to be heard in order to demonstrate how "Devastation" (open-cast lignite mining) affected the cultural heritage of the region (www.forst-lausitz.de/sixcms/media.php/674/Broschuere_AVO_Aufl2.pdf).

Jodlbauer, R., Spieß, G. & Steenwijk, H. (2001). *Die aktuelle Situation der niedersorbischen Sprache: Ergebnisse einer soziolinguistischen Untersuchung der Jahre 1993-1995*. Bautzen: Domowina (= Schriften des Sorbischen Instituts 27).

Muka, Ernst (1911-1926). *Słownik dolnoserbskeje rěcy a jeje narěcow I–III*. Petrograd: RAN; Praha: ČAVU.

Norberg, Madlena (1996). *Sprachwechselprozeß in der Niederlausitz. Soziolinguistische Fallstudie der deutsch-sorbischen Gemeinde Drachhausen/ Hochoza*. Uppsala (= Acta Universitatis Upsaliensis. Studia Slavica Upsaliensia 37).

Scholze, Lenka (2008). *Das grammatische System der obersorbischen Umgangssprache im Sprachkontakt.* Bautzen: Domowina (= Schriften des Sorbischen Instituts 45).

Schwela, Gotthold (1906). *Lehrbuch der Niederwendischen Sprache. Erster Teil: Grammatik*. Heidelberg: Ficker.

SSA 1-15 1965-1996 *Sorbischer Sprachatlas* (Serbski rěčny atlas), bearbeitet von H. Faßke, H. Jentsch und S. Michalk, 1-15, Bautzen (Budyšin) 1965-1996.

Starosta, Manfred (1991). *Niedersorbisch schnell und intensiv 1*. Bautzen: Domowina.

Sundara, M., Polka, L., & Baum, S. (2006). Production of coronal stops by simultaneously bilingual adults. *Bilingualism: Language and Cognition* 9, pp. 97–114.

Internet sources (accessed 30.03.2011):

www.forst-lausitz.de/sixcms/media.php/674/Broschuere_AVO_Aufl2.pdf

http://genie.coli.uni-saarland.de

http://www.mpi.nl/DOBES/

http://www.eva.mpg.de/lingua/resources/lela.php

http://endangeredlanguagefund.org/