This is the published version of a chapter published in *Språkdidaktik: Researching Language Teaching and Learning*.

Citation for the original published chapter:

Developing a spoken corpus for South Saami language teaching and learning.
In: Eva Lindgren, Janet Enever (ed.), *Språkdidaktik: Researching Language Teaching and Learning* (pp. 75-84). Umeå: Department of Language Studies, Umeå University

N.B. When citing this work, cite the original published chapter.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-52710
DEVELOPING A SPOKEN CORPUS FOR SOUTH SAAMI LANGUAGE TEACHING AND LEARNING

Mikael Vinka, Christian Waldmann, David Kroik & Kirk P H Sullivan

Introduction
Over the past 20 years Corpus Linguistics has increasingly been used to support language teaching and learning by providing exposure to authentic language (Chambers, 2007). In the context of minority and indigenous language teaching and learning, Boyce (2006) argues that modern spoken language corpora provide a means of overcoming limited language exposure for both the first and second language speaker. Granger, Kraif, Ponton, Antoniadis and Zampa (2007) pointed out “native corpora give no indication of what is difficult for learners” (p. 253) suggesting a need for learner corpora as a complement to native speaker corpora and that learner corpora have the potential to support curriculum design for second language education, and to inform language revitalization initiatives and educational practices. In this chapter we discuss and illustrate the issues associated with developing a spoken first and second language corpus for the Scandinavian minority language South Saami, and we suggest ways in which it could be used to support South Saami language education.

To set the scene, we begin this chapter by overviewing the Saami languages and specifically the South Saami situation to provide the reader with an understanding of the context in which we are developing a modern spoken language corpus, before overviewing the on-going Saami corpus development projects that have explicit educational objectives. Thereafter, we consider the specific challenges the South Saami context poses for the development of a spoken South Saami language corpus. In the final part of this chapter, we discuss how this corpus could be used to inform language teaching and learning.

Background
The Saami languages constitute a branch of the Finno-Ugric languages. There are nine living Saami languages according to UNESCO (http://www.unesco.org/culture/languages-atlas/en/atlasmap.html). Three (Ter, Pite, and Ume) are critically endangered with circa 10, 20, and 20 speakers, respectively. Five (Kildin, Skolt, Inari, Lule and South) are severely endangered with circa 800, 300, 400, 2000 and 500 speakers, respectively. One (North) is defined as definitely endangered with circa 30 000 speakers.

Although the Saami languages are closely related and commonly perceived to be mutually intelligible, they differ in several non-trivial linguistic aspects. For example, the phonological, morphological, lexical and syntactic distinctions between North and South Saami are so sharp (e.g. Sammallähti 1998; Vinka in press) that these languages cannot be considered minor mutually intelligible regional variants of each other (see example 1 and 2).
South Saami is spoken in central Norway and Sweden across an area that stretches approximately 600 kilometres from north to south and 300 kilometres from east to west. Rasmussen and Nolan (2011) estimated that the language has up to 1000 speakers of which half may be second language learners. This results in a low speaker density, which aggravates the frail situation of the language (Grenoble & Whaley 2006).

Although the future of South Saami may seem bleak, there are indications that the intergenerational transfer has developed benevolently (Sametinget 2012). This development is a consequence of a strong and active language revitalization movement that has emerged over the past two decades (Todal 2007). In lieu of these positive tendencies, the language is extremely frail, and in order to ensure that coming generations speak South Saami, massive efforts are required in several areas.

South Saami gained legal recognition in Sweden in 2009, when the Bill on National Minorities and Minority Languages was passed (SFS 2009:724). This law grants among other things, the right for Saami children to attend Saami-speaking kindergarten. However, this law is silent on matters pertaining to the formal educational system, which is regulated by Skollagen (SFS 2010:800) and Skolförordningen (SFS 2011:185). According to Skollagen, Saami children have the right to receive language classes. However, the Law only specifies that such classes are extra-curricular, and leaves it open to local authorities to decide whether or not to grant classes (for instance SFS 2010:800 10Kap, §7). Furthermore, the legal minimum language exposure is 45 minutes per week, which is not sufficient to achieve proficiency in a language (Svonni 1993, 2008).

Matters are further aggravated by Skolförordningen (SFS 2011:185) which in effect prohibits the establishment of revitalization programs based on full and partial immersion methods. Skolförordningen (SFS 2011:185 9Kap, 12§) states that language of instruction may be a language other than Swedish up to 50 percent of the instruction. This explicitly bans full immersion which have proven to be extremely successful (Grenoble & Whaley 2006). While bilingual instruction may take place, it is however conditioned by the requirement that it takes place during a transition period that results in 100 percent instruction in Swedish (SFS 2011:185 9Kap, 12§). Hence long-term partial immersion is also illegal. In short, the Swedish educational system is governed by laws that enforce an explicit assimilation policy, and thus stands in sharp contrast to Norway and Finland, where immersion models in the formal educational system are legal (Finlands regering 1998; Justisdepartementet 1998).

The everyday use of South Saami is limited even in the core area with the language mostly spoken within the homes of South Saami families. The use in the public domain is in essence non-existent. The spoken South Saami in the one radio programme broadcast each week in South Saami is, adjusting for musical content, approximately 6–7 minutes long.
Given the context we have described, there are in addition to the challenges faced by all minority languages specific challenges for South Saami language teaching and learning. One such challenge is the shortage of teaching materials such as authentic texts and recordings, and linguistically informed dictionaries and grammar books. Such materials can be developed using a corpus of spoken language, which can be used to develop linguistically informed teaching and learning materials based on authentic contemporary language use and pronunciation. The challenges associated with the South Saami situation affect the process of constructing a corpus of spoken South Saami, and delimit the ways in which such a corpus can and should inform the development of new language teaching and learning materials. We illuminate these challenges by first considering on-going North and Lule Saami spoken language corpus development projects that have explicit educational objectives before turning to the specific challenges facing the development and use of a corpus of contemporary spoken South Saami.

Saami language corpora

In the Saami context, there are currently a few on-going projects developing written and spoken language corpora with explicit educational objectives. The written language corpora being developed at the Centre for Saami language technology (Giellatekno) at Tromsø University (http://giellatekno.uit.no/) together with Divvun (a project and working group initiated by the Norwegian Saami Parliament in 2005 http://divvun.no/) are the most well known. Giellatekno is working on building written corpora for several of the Saami languages. The corpora have, among other things, been used to develop computer-based tools for supporting the languages (e.g. machine translation, spellcheckers, and digital dictionaries). One example of a pedagogical tool that has been developed at Giellatekno (in cooperation with Aajege and Gielejarnge, two language centres located in Røros in Norway and Östersund/Tärnaby in Sweden) is OAHPA!, a language learning program for North Saami and South Saami. Its focus is on grammar and vocabulary and it is particularly useful for second language learners regardless of age and language proficiency. OAHPA! can also be used by first language learners, for example, to develop writing and vocabulary skills.

One current major spoken language corpus project for the Saami languages is DASAGO, Davvisámí mánáid giellaovdáneapmi (‘North Sami child language acquisition’) that is developing a corpus of contemporary spoken North Saami (https://castl.uit.no/index.php/acquisition/dasago). One part of the project is to document mono- and bilingual first language acquisition of North Saami, and to develop a corpus of spontaneous North Saami child language acquisition and adult child-directed speech. The child-directed speech will provide valuable information about the quality and quantity of North Saami input that the children receive. This is important for understanding how mono- and bilingual children learn North Saami. Outside the North Saami core area there is large variation with respect to the amount of North Saami input that children receive; most children acquire North Saami and at least one other national language (Finnish, Norwegian, Russian, and/or Swedish). Children who receive a limited North Saami input and have Saami as their weaker language are at a risk of language delay in North Saami. Høier (2007) showed that children with a
Saami language as their weaker language may face difficulties when learning to read their Saami language.

However, with appropriate, informed linguistic training input-related language delays can be overcome. This requires an understanding of the monolingual as well as bilingual development of modern spoken North Saami, something that the DASAGO corpus will support. The DASAGO project aims to develop language pedagogical strategies to compensate for the scarcity of North Saami input and to better support the language development of children with Saami as their weaker language. Such strategies will be useful for day-care centres and for schools.

In this overview of existing and under-construction Saami Language Corpora we have demonstrated how it is possible to create corpora for indigenous languages and that these corpus projects have language teaching and learning applications that support the maintenance and revitalization of the languages. In the remainder of this chapter we consider the development of a spoken language corpus for South Saami that is a minority language with fewer speakers and that therefore poses greater challenges for corpus construction and use in the language teaching and learning context.

The creation of a corpus on spoken South Saami
The process of creating spoken language corpora that are of relevance for education is complex. Drawing on the experiences of creating the abovementioned corpora, we explore the complexities of spoken minority language corpus creation through an ongoing project. Building a corpus of spoken language involves procedures such as collecting spontaneous speech and transcribing it. These procedures need to be performed in such a way that the corpus becomes usable for the intended user. The corpus has educational objectives and will be developed to inform South Saami language teaching and learning at all language proficiency levels across the entire educational system, and to provide a base for linguistic research into the language and its language revitalization. Educationally, it is important to have knowledge about the target language as well as about the learner language. To provide a picture of both the target language and the learner language, the corpus will include spontaneous speech from strong first language speakers (L1 speakers) as well as from second language learners (L2 speakers).

In the South Saami context, the low number of speakers, the even lower number of L1 speakers, and the weak intergenerational transfer of the language pose major difficulties for the collection of spoken language. Today, many South Saami speakers are L2 speakers and have learnt the language in school settings. Swedish, Norwegian, and/or North Saami influence these speakers’ South Saami in other ways than these languages influence the L1 speaker’s South Saami. As Mæhlum (2007) pointed out all linguistic aspects of South Saami (pronunciation, lexicon, morphology, syntax) have been affected by the national languages, Swedish and Norwegian, and North Saami.
For language teaching and learning it is important to understand the language learning process. We, therefore, acknowledge following Granger et al. (2007), the contribution that speech from L2 learners will make to a corpus on spoken South Saami. Spontaneous speech from L2 learners is an important resource for understanding the process of learning South Saami. The corpus will reveal, for example, typical errors, learning paths, learning strategies, and transfer effects. The revealed features will play an important role in developing learner-corpus-informed language teaching and learning tools and materials such as “dictionaries, grammars, textbooks, CALL programs — that address learners’ attested difficulties” (p. 253) of relevance for educational and pedagogical strategies in schools and universities.

These L2 difficulties illustrate why it is important to find strong L1 users and utilize their native intuitions about contemporary South Saami when developing teaching and learning materials. However, the scarcity of such strong L1 users makes collecting authentic L1 speech difficult. The problem identifying strong L1 users also creates difficulties for producing descriptions of the linguistic aspects of the contemporary target language that are important for designing relevant teaching and learning materials. The accuracy of these descriptions is important as inaccurate linguistic descriptions can misinform educational and revitalization policies and strategies, and result in incorrect teaching and learning materials, pedagogical strategies, course contents and the spread of incorrect knowledge in and about the language.

Although collecting spontaneous speech from strong L1 speakers poses a difficulty in the case of South Saami, the collection procedure is only a minor phase in building a corpus of spoken language. The most time-consuming phase in building a spoken language corpus is the transcribing and tagging of the audio recordings. The norms of the written language present an apparent bias when transcribing speech. A known pitfall is that transcribers tend to map the language form of the speaker onto a standard written form when there is no real correspondence (overnormalization). The less proficient the transcriber is in the language, the greater the risk is for such overnormalization.

For example, if a speaker says *ajja tuhtj* ‘grandfather thinks’, a transcriber may instead transcribe the standard written form *aajja tuhtjie* ‘grandfather thinks’. The transcriber’s overnormalizing misses valuable information about a common phonological deletion in spontaneous spoken South Saami. The transcriber fails to note the deletion of the word final -ie in *tuhtjie*, that occurs in sixth conjugation verbs (Bergsland 1994). Further, the transcriber misses that a stressed vowel that appears in a closed syllable shortens in the spoken language (*a*- in *ajja*), even though it is spelt as if it were a long vowel (*aa*- in *aajja*).

It is paramount that these inaccurate transcriptions do not occur because knowledge about common phonological deletions/reductions in spontaneous speech, and typical learner errors and learning paths are valuable when designing curricula for language courses and when constructing teaching and learning materials.

To deal with the bias of the written language norms when transforming speech into written text it is important that the transcribing process is guided by standardized transcription conventions. One such format is CHAT (Codes for the Human Analysis of Transcripts) that
is a standardized transcription and coding scheme developed as part of the CHILDES (Child Language Data Exchange System) project (MacWhinney 2000; http://childes.psy.cmu.edu/). CHAT was designed to increase the reliability of transcriptions, and it is widely used in child language acquisition research, and has also been used in the development of spoken learner corpora such as the French Learner Language Oral Corpora (Myles & Mitchell, http://www.flloc.soton.ac.uk/) and Spanish Learner Language Oral Corpora (http://www.sploc.soton.ac.uk). CHAT allows phonetic transcription using the International Phonetic Alphabet (IPA), the indication of nonstandard lexical forms, as well as the linking of audio to the transcripts. As South Saami is an understudied language and little is written about the phonological processes in spontaneous speech, there is a pedagogical benefit in including phonetic transcriptions in the corpus. This will highlight the differences between written and spoken South Saami, and contribute to the understanding and teaching of pronunciation in fluent speech in a way that existing teaching and learning materials cannot.

CHAT is fully compatible with a range of programs such as Phon (http://www.phon.ca/) and Praat (www.fon.hum.uva.nl/praat/). Phon is particularly valuable for the phonetic transcribing of spontaneous speech, as it enables multiple blind transcriptions, which can be compared to the original sound source, and thus eliminate many of the pitfalls that accompany the transcription process. Both Phon and Praat allow users to conduct phonological and phonetic analyses of data transcribed in CHAT. Further, the CHAT system also includes CLAN (Computerized Language ANalysis) programs, which provides tools for a wide range of automatic searches and detailed phonological, morphological and syntactic analyses of transcripts in the CHAT format. Such tools can be useful not only for teachers but also for students at all levels when using a corpus on spoken South Saami.

Discussion and conclusions
The importance of developing a corpus of spoken South Saami that will inform language teaching and learning becomes apparent when summarizing the language situation in the following eight points:

1. Few South Saami speakers/learners are as competent in South Saami as in the national languages

2. Those who use South Saami daily have problems raising their proficiency level

3. South Saami speakers/learners have to actively search for opportunities to use the language

4. There is little literature for speakers/learners to read in South Saami

5. The range of South Saami medium television and radio programmes verges on the non-existent
6. The growth and strength of their national language competence is supported in everyday life.

7. The growth and strength of their South Saami is not supported in everyday life.

8. The speakers/learners need to maintain their commitment to South Saami to maintain and improve their written and spoken language proficiency.

Carter and McCarthy (1988) among others have pointed out that language learners ought to learn the language as it is actually used by native speakers. It is thus important that any modern spoken corpus is used to test existing descriptions against authentic current language use and assists in the development of state-of-the-art linguistic descriptions. These descriptions can counter the overconfidence some language users have in standard grammars and pronunciation rules that encourage them to change their way of talking away from the contemporary to what they believe to be “correct” as stated in grammar and pronunciation guides that do not describe contemporary language use. Such descriptions can readily be tested against a modern spoken language corpus of authentic data, and revised descriptions can inform the development of contemporary teaching and learning materials. Further in a teaching and learning context, all descriptions can easily be illustrated with authentic modern examples taken from the corpus.

Language learner pronunciation is frequently affected by orthographic representations and spelling rules. Although this is a general problem in many language learning situations, e.g. the classroom, the poverty of readily available authentic language input for the South Saami learner increases the impact of orthographic representation on pronunciation and, in particular, the learning of phonological processes. For example, the South Saami word *rööpes* 'red' /røps/ is frequently incorrectly pronounced by language learners as /rø:psVs/. This is caused by pronouncing all graphemes in the written word including the ‘c’ and by following the taught South Saami spelling rule that a double written vowel is always pronounced as a long vowel resulting in /øː/ rather than the short vowel /ø/. In the absence of access to authentic modern spoken input and the near-absence of South Saami in radio and TV news programmes or other media a modern spoken corpus can help this situation in at least two ways. One, as is often the case, spelling rules are not transparent and a simple spelling rule can often obscure a more complex rule. This is evidenced in our preliminary corpus of South Saami in which it is clear that long vowels only occur in stressed open syllables. Hence, the teacher and the learner need to be aware that the simple rule that a double written vowel represents a long vowel needs to be modified to the more complex rule: a double written vowel represents a long vowel only when the syllable is stressed and open. Two, having access to authentic modern spoken South Saami recordings can support the learner’s acquisition of pronunciation and reduce the dependence on orthographic representation.

Access to authentic spoken material also provides the teacher and the learner opportunities to work at linguistic levels above the word. A learner with advanced grammatical and lexical knowledge, yet little advanced interaction with native speakers can lack knowledge about discourse particles and their uses. A modern spoken corpus would provide valuable insights...
into the use of discourse particles in contemporary South Saami. Together this will assist in improving authenticity of teaching and learning for natural interaction.

In our work, we actively use the recorded data for state-of-the-art linguistic descriptions to inform teaching and learning materials for the lexicon and beyond, and to provide authentic examples of the spoken language. We anticipate that the biggest contribution of a modern spoken South Saami language corpus is attitudinal. The teaching of the language is a core element of language revitalization and this can be supported by having a state-of-art linguistic description of the language, informed teaching materials, and contemporary authentic examples of the spoken language from both first and second language speakers of the language. The resultant teaching and learning materials we anticipate will make the language learning experience effective and enjoyable and thus produce strong advocates for revitalization.

References


URL LIST
http://giellatekno.uit.no
http://divvun.no/
https://castl.uit.no/index.php/acquisition/dasago

http://childes.psy.cmu.edu/

http://www.flloc.soton.ac.uk/

http://www.sploc.soton.ac.uk

http://www.phon.ca

www.fon.hum.uva.nl/praat/