Handwritten Text Recognition and Linguistic Research

Erik Magnusson Petzell

Institute for Language and Folklore, Vallgatan 22, S-41116, Gothenburg, Sweden
erik.petzell@isof.se

Abstract. This paper presents ongoing work with automatic transcription of handwritten, phonetically precise dialect texts from the south-west of Sweden (collected in the 1890s). Using a SAMPA-based transcription key (where SAMPA stands for Speech Assessment Methods Alphabet), I have enabled the training of an HTR engine (where HTR stands for Handwritten text recognition), by feeding it manual transcriptions, to automatically transcribe two separate (but similar) phonetic hands. The phonetically detailed output reveals structural properties of the dialect that are hard (at best) or impossible (at worst) to retrieve from other sources. In this paper, I show how my research on enclitic pronouns in North Germanic has benefitted from the possibility to search for prosodic dependencies that the digital versions of the dialect texts provide.

Keywords: Handwritten Text Recognition, Dialect Texts, Swedish, International Phonetic Alphabet, Speech Assessment Methods Alphabet, Digitization, Enclitic Pronouns.

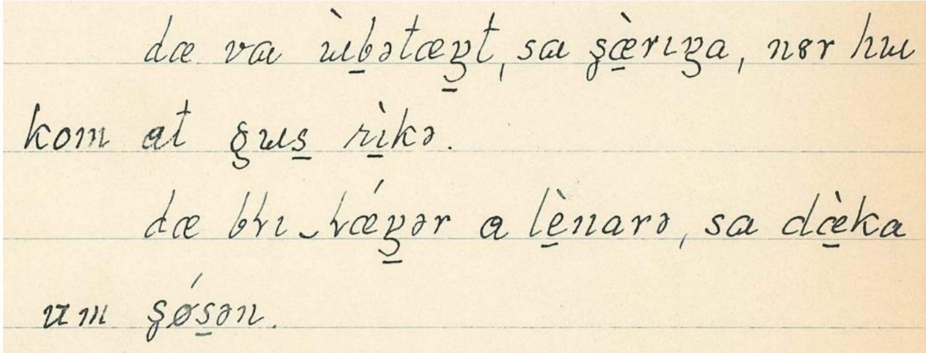
1 Introduction

In this paper, I describe my ongoing work with automatic transcription of handwritten Swedish dialect texts from the 19th century, and relate it to my linguistic research. I specialize in morphosyntactic variation during earlier stages of North Germanic, as manifested both in historical texts and in archaic dialects. In recent years, an important domain of inquiry for me has been enclisis [5, 6], which is a type of linguistic phenomenon that balances on the border, as it were, between syntax and morphology. For instance, enclitic pronouns fill syntactic slots just like free pronouns and larger noun phrases. However, clitics are prosodically dependent on another word, in effect being unable to bear stress. In that respect, they are more like inflectional endings than independent phrases.

Enclisis of any kind is hard to investigate in texts, since orthography, both in the past and the present, normally does not mark it. Audio recordings of dialect speakers may contain relevant data for historical linguists, but this type of material is very time consuming to work with. There is, at present, no working method for retrieving linguistic structure from an audio file containing non-standard language [1]. Certainly, the Text laboratory at Oslo University harbours a large corpus of manual transcriptions of dialect audio from all over Scandinavia: The Nordic Dialect Corpus [3]. This corpus is a great tool for investigating for instance word order variation, but the phonetic details of the recordings are only very rudimentarily rendered by the transcript. Enclitic status, for instance, is not marked at all.[1] Furthermore, audio recordings go only so far back in time, for obvious technical reasons. Although we have sporadic field recordings already from the 1930s, it is not until the late 1940s that this form of documentation starts generating the great bulk of dialect audio still kept at the Scandinavian dialect archives.

2 Phonetically Precise Dialect Texts

There is a third type of archival language data, which constitutes an intriguing source of linguistic structure of old: dialect texts, handwritten in the 19th century using a traditional phonetic alphabet. Such texts exist in archives all over Scandinavia, and through them, we are granted access to the phonetic subtleties of an era that is too distant to have been caught on audio tape. So far, I have only scraped the surface of this great pile of detailed dialect data. For practical reasons, I have started with texts from the dialect archive in Gothenburg, where I work.[2] I will refer to the alphabet used in these texts as LMA, a label based on the name of the Swedish dialect alphabet (viz. LandsMålsAlfabetet, 'the alphabet for rural dialects'). Figure 1 below shows four lines of LMA from a compilation of dialectal expressions, collected in the parish of Fagered (in the south-west of Sweden) in the beginning of the 1890s; below the image is a word for word translation into Standard Swedish and English.



dæ va ù̱bətæɣ̱t, sa ʑæ̱̀riɣa, nɛr hu
kom at ɡus̱ rìḵə.
dæ bli‿ɹǽɣ̱ər a lè̱narə, sa dæ̱̀ka
ɯm ʑǿs̱ən.

Fig. 1. Four lines from the Fagered collection.

det var obetänkt, sa käringen, när hon (Stand. Sw.)
kom till Guds rike.
det blir längre och lenare, sa flickan
om kyssen.

[1] See p. 13 in the transcription guidelines: http://www.tekstlab.uio.no/nota/scandiasyn/Transkripsjonsrettleiing%20for%20ScanDiaSyn.pdf

[2] This archive is part of a national government agency, The Institute for Language and Folklore, forming a Department of Dialectology, Onomastics and Folklore Research. See: http://www.isof.se/om-oss/verksamhet/about-the-institute.html for general information in English.

‘it was inconsiderate, said the-woman, when she (Eng.)
came to God’s kingdom.
it becomes longer and leaner, said the-girl
about the-kiss.’

Texts of this sort should not be mistaken for straightforward transcripts of spontaneous speech. Instead, we can think of them as prototypes of the parish dialect of the time, created by the field linguist and based on his encounters with several local informants over a longer period of time. As it stands, the font encodes many aspects of the acoustic signal: it conveys the precise phonetic value of each segment, it marks length within stressed syllables (underscoring the long segment), as well as the melodic characteristics of all disyllabic (and longer) words, distinguishing between acute, ´, and grave, `, accent (which is a phonologically relevant difference in many Scandinavian dialects). Finally, it also marks prosodic dependence between words, tying them together with a bow-shaped linking sign (appearing between word 2 and 3 on the third line in the image). As stated above, prosodic dependence is characteristic of enclisis.

Nowadays, the LMA is used very marginally (and almost never outside of traditional onomastics). As a rule, linguists of today instead use the International Phonetic Alphabet,[3] when there is need for phonetic detail in written form. However, as soon as corpus-based linguistic research targets non-phonological issues, the fine phonetic details are superfluous. In fact, such detail only makes word- and phrase-based searches more complicated. Consequently, in order to make the old dialect texts useful for different sorts of linguistic research, it does not suffice to simply transform the text of the images to a digital correlate. In addition, there is need for several conversions of the original text into different more or less simplified formats, which, in turn, can be useful also for non-linguists (both other researchers and members of the general public).

3 From Handwritten Phonetics to Searchable Text

In this section, I present my initial attempts to transfer the old handwritten Swedish dialect texts into a digital format. This work started last year within the project Tilltal [1], but as of this year, it is also conducted within the infrastructure The National Language Bank.[4]

The tool I use to analyse and transcribe the texts is Transkribus.[5] As a first practical step, I had to decide how to write the LMA with my standard keyboard. As mentioned, the LMA is hardly used anymore, and only very few of the LMA symbols have a Unicode status. Although all IPA symbols indeed do, they are difficult to produce with a standard keyboard. In order to reach an acceptable transcription speed, I have instead created a SAMPA-based transcription key. SAMPA stands for Speech Assessment Methods Phonetic Alphabet and it resorts only to the 128 characters that a standard (i.e.

[3] See https://www.internationalphoneticassociation.org.

[4] See http://www.sprakbanken.se/eng.

[5] See https://transkribus.eu/Transkribus/.

English) keyboard can produce.[6] These characters, either in isolation or combined with others, are then given a specific phonetic value. Although the underlying principles for creating phonetic symbols are the same, my dialect SAMPA is a digital version of the LMA and is therefore quite different from standard Swedish SAMPA, which is IPA-based.[7]

To begin with, I made a SAMPA transcript of roughly 100 pages of the Fagered collection, from where the example in Figure 1 is drawn. This amount of manual transcription is what is needed to train a so called HTR engine (where HTR stands for handwritten text recognition). Once the HTR engine is integrated in the Transkribus platform, it is capable of automatically generate transcriptions of more text of the same hand. How well the engine works of course depends on an array of factors. One factor that often (according to the Transkribus crew) turns out to be complicating is super- and subscripted diacritics of the sort that occur abundantly in the dialect texts. Still, the HTR engine managed to handle the rest of the Fagered collection almost flawlessly; only a handful of minor manual corrections (concerning individual segments or diacritics) per page (16 lines) was required to perfect the transcription. Figure 2 below shows the automatic transcription of the four lines given above (here, the automatic transcription was flawless to begin with). As can be seen, each LMA segment has its SAMPA counterpart, the accents correspond to single (acute) or double (grave) quotation signs, and length and prosodic dependence are marked by colon and underscore respectively. For a word for word translation, see Figure 1 above. For a complete list of the correspondences between LMA and dialect SAMPA, see the appendix.

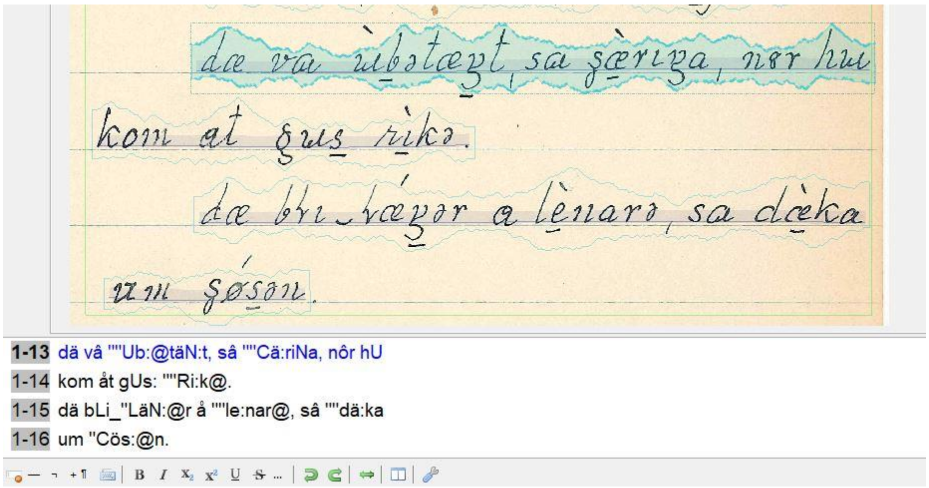
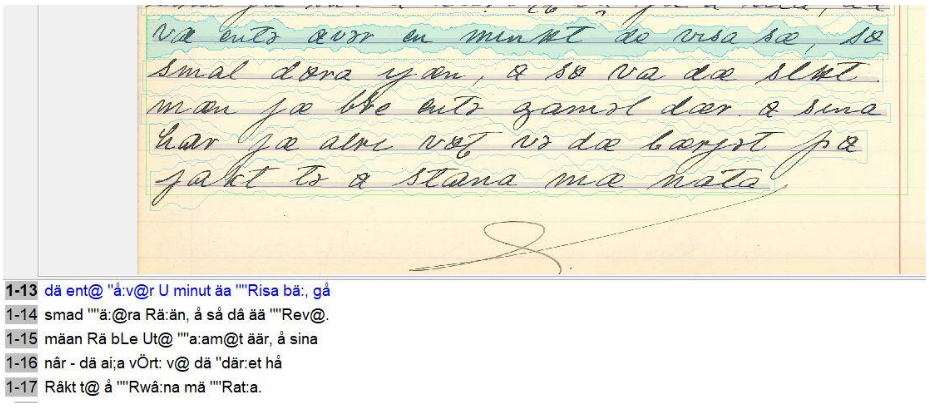

1-13 dä vâ ""Ub:@täN:t, sâ ""Cä:riNa, nôr hU
1-14 kom åt gUs: ""Ri:k@.
1-15 dä bLi_"LäN:@r å ""le:nar@, sâ ""dä:ka
1-16 um "Cös:@n.

Fig 2. Automatic transcription (Fagered collection).

[6] See https://www.phon.ucl.ac.uk/home/sampa/.

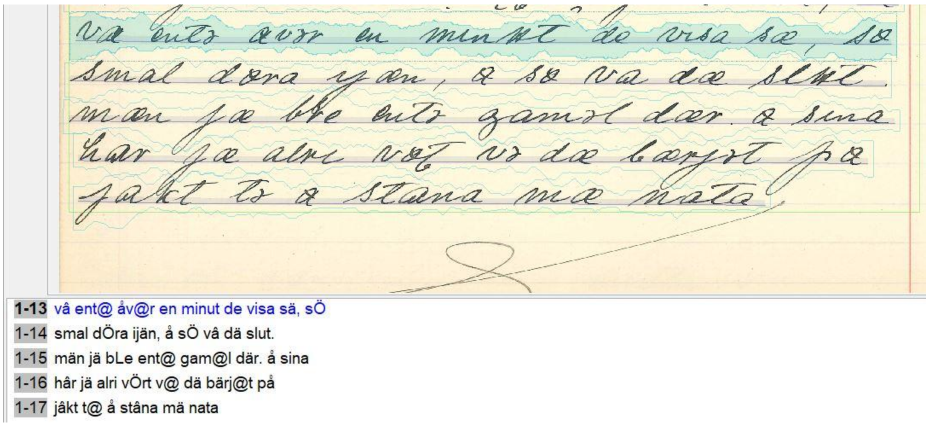
[7] See https://www.phon.ucl.ac.uk/home/sampa/swedish.htm. A complete account of the differences between Standard Swedish SAMPA and dialect SAMPA can be found in the appendix.

Transcription accuracy naturally decreases dramatically when the HTR engine is run on other LMA texts, written by other field linguists. Figures 3 and 4 below show five lines from a text from about the same time as the Fagered collection, documenting the contemporary dialect of the island of Orust in the province of Bohuslän. Figure 3 shows what happens when the Fagered engine tries to handle text from Orust. Only about a third of the LMA words are represented correctly in the SAMPA format. However, by adding some 50 pages of manual transcription of Orust text to the training sample of the existing HTR engine, the resulting SAMPA output becomes as satisfactory as with the Fagered collection; see Figure 4 and the word for word translation below. As an indication of HTR precision, the page from where these five lines are drawn (1 page = 16 lines as before) required 5 manual corrections: the substitution of one ŋ for ŋ̧, three deleted incorrect commas, and one added full stop.



1-13 dä ent@ "å:v@r U minut äa ""Risa bä:, gå
1-14 smad ""ä:@ra Rä:än, å så dâ ää ""Rev@.
1-15 mäan Rä bLe Ut@ ""a:am@t äär, å sina
1-16 nâr - dä ai;a vÖrt: v@ dä "där:et hå
1-17 Râkt t@ å ""Rwâ:na mä ""Rat:a.

Fig. 3. Automatic transcription (Orust collection, Fagered engine).



1-13 vâ ent@ åv@r en minut de visa sä, sÖ
1-14 smal dÖra ijän, å sÖ vâ dä slut.
1-15 män jä bLe ent@ gam@l där. å sina
1-16 hâr jä alri vÖrt v@ dä bärj@t på
1-17 jâkt t@ å stâna mä nata

Fig. 4. Automatic transcription (Orust collection, modified engine).

var inte över en minut det visade sig, så (Stand. Sw.)
small dörren igen, och så var det slut
men jag blev inte gammal där. och sedan

har jag aldrig varit vid det berget på
jakt till att stanna med natten
‘was not over a minute it showed itself, then (Eng.)
shut the-door close, and so was it over
but I became not old there. and then
have I never been by that the-mountain on
hunt in to stay with the-night’

4 Future Tasks

Apart from dealing with the actual transference process (i.e. LMA image → SAMPA transcript), I have also experimented with conversions from SAMPA to other more or less simplified formats, in order to make the texts accessible for a wider circle of users (see [7]:196–197 for examples). Only quite recently have I become aware of the models for dialect transliteration developed by the Text Laboratory in Oslo.[8] These models transform dialectal forms to standard language, which opens up for automatic lemmatization and annotation, in turn enhancing searchability radically. My ambition is to learn from the Norwegian project and to add transliteration to standard Swedish to the list of formats that the SAMPA transcripts can be converted to.

Finally, I will make a brief note on how my linguistic research into clitics has been facilitated by the digitization of dialect texts. Since the SAMPA output contains both phonetic and prosodic details, it is fairly easy to extract those instances of prosodic dependencies (marked _ in the SAMPA format) in a text that represent potential enclitic pronouns. A somewhat prosaic effect hereof is simply that I am now able to sort and quantify relevant data in a way that I could not do before. A more intriguing consequence is that I have actually discovered linguistic variation that has previously fallen under the radar. For instance, in his description of traditional Bohuslän dialects, Kvillerud mentions only one masculine and one feminine object clitic*: (e)n* and *(n)a* respectively ([4]: 95), both special clitics in the sense of Zwicky ([8]; see also [2]: 28–31). However, my Orust text reveals a hitherto unnoticed gender asymmetry: the feminine form *(n)a* in fact competes with a simple clitic, *ner*, which is a reduced form of the full pronoun *hener*, whereas *(e)n* remains the only masculine option, reduced forms of the full pronoun *ham* being unattested. How to understand this asymmetry is an exciting question for future research.

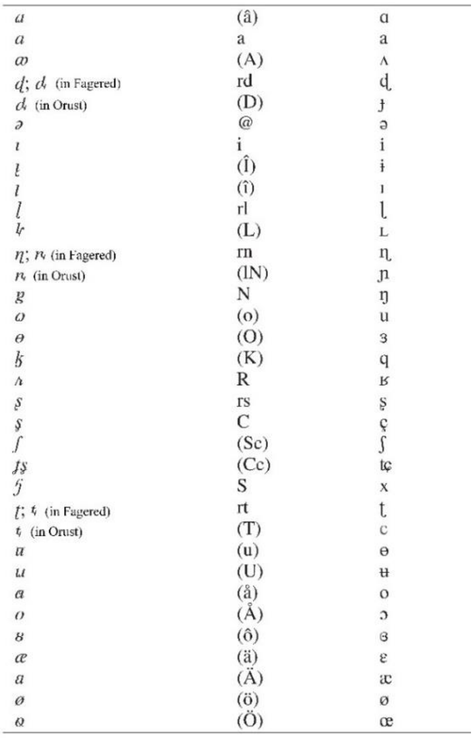
Appendix

Correspondences between LMA (column 1), dialect SAMPA (column 2) and IPA (column 3) are shown in table 1 (individual segments) and table 2 (diacritics) below. In table 2, *e* represents a segment, and *s* represents a sequence of segments separated by space or _. SAMPA symbols within parentheses have been created by me. The remaining SAMPA symbols are Standard Swedish SAMPA. Some symbols (<b>, <e>, <f>,

[8] See http://tekstlab.uio.no/LIA/pdf/rettleiing-translitterator.pdf.

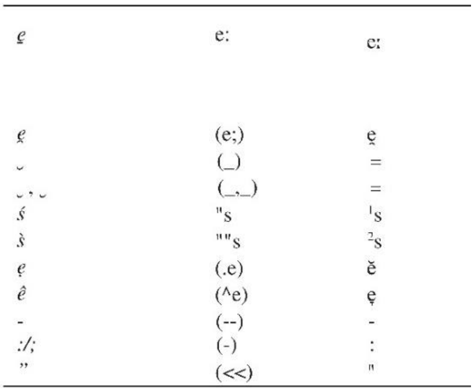
<g>, <h>, <j>, <k>, <l>, <m>, <n>, <p>, <r>, <s>, <t>, <v>, <y>) and diacritics (‹,›, ‹.›, ‹!›, ‹?›, ‹(›,‹)› are the same in the LMA original and my SAMPA transcript, and are therefore not included in the tables.

Table 1. Segments.



ɑ	(â)	ɑ
a	a	a
ꭤ	(A)	ʌ
ɖ; d̢ (in Fagered)	rd	ɖ
d̢ (in Orust)	(D)	ɟ
ə	@	ə
ι	i	i
ι̵	(Î)	ɨ
l	(î)	ɪ
ɭ	rl	ɭ
ɫ	(L)	ʟ
ɳ; n̢ (in Fagered)	rn	ɳ
n̢ (in Orust)	(lN)	ɲ
ŋ	N	ŋ
ω	(o)	u
ɵ	(O)	ɜ
ꝁ	(K)	q
ʌ	R	ʁ
ʂ	rs	ʂ
ş	C	ç
ʃ	(Sc)	ʃ
ʧ	(Cc)	tç
ɧ	S	x
ʈ; t̢ (in Fagered)	rt	ʈ
t̢ (in Orust)	(T)	c
u	(u)	ɵ
ɯ	(U)	ʉ
α	(å)	o
ο	(Å)	ɔ
ȣ	(ô)	ɞ
æ	(ä)	ɛ
a	(Ä)	æ
ø	(ö)	ø
ꝋ	(Ö)	œ

Table 2. Diacritics.



ȩ	e:	eː
ḝ	(e;)	e̝
͜	(_)	=
͜ , ͜	(_,_)	=
ś	"s	ˈs
s̀	""s	²s
ẹ	(.e)	ĕ
ê	(^e)	e̞
-	(--)	-
:/;	(-)	ː
”	(<<)	"

References

1. Berg, J., Domeij, R., Edlund, J., Eriksson, G., Fallgren, P., House, D., Lindström, E., Petzell, E. M., Malisz, Z., Nylund Skog, S., Öqvist, J.: Making archival speech recordings accessible for research. Svenska landsmål och svenskt folkliv 141, pp. 171–178 (2019), http://gustavadolfsakademien.se/files/download/documents/SvLm2018.pdf
2. Howe, S.: The personal pronouns in the Germanic languages: a study of personal pronoun morphology and change in the Germanic languages from the first records to the present day. de Gruyter, Berlin (1996).
3. Johannessen, J. B., Priestley, J., Hagen, K., Åfarli, T. A., Vangsnes, Ø. A.: The Nordic Dialect Corpus – an Advanced Research Tool. In: Jokinen, K., Eckhard, B. (eds.): Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4 (2009), https://www.aclweb.org/anthology/W09-4600.
4. Kvillerud, R.: Bohuslänska. Språkprov med kommentar. Skrifter utgivna av Språk- och folkminnesinstitutet, Dialekt-, ortnamns- och folkminnesarkivet i Göteborg 5. Språk- och folkminnesinstitutet, Gothenburg (1999).
5. Petzell, E. M: Enclitic subjects and agreement inflection in Viskadalian Swedish. Nordic Atlas of Language Structures (NALS) Journal 2, 1–39 (2017), http://dx.doi.org/10.5617/nals.5360
6. Petzell, Erik M.: Head conjuncts: evidence from Old Swedish. Linguistic Inquiry 48(1), 129–157 (2017b), https://www.mitpressjournals.org/doi/pdf/10.1162/LING_a_00237.
7. Petzell, E. M.: Automatisk transkribering av landsmålstext. Svenska landsmål och svenskt folkliv 141, 184–199 (2019), http://gustavadolfsakademien.se/files/download/documents/SvLm2018.pdf
8. Zwicky, A. M.: Clitics and Particles. Language 61, 283–305 (1985).

Archival sources

1. The Fagered collection = Accession number DAGF 269F: I–II
2. The Orust collection = Dialect texts IOD, old accession numbers: 22:1–3, 27:1–9.