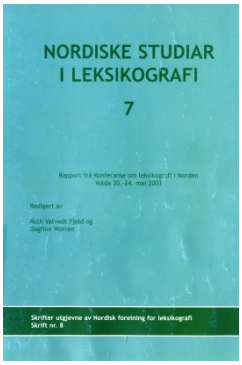
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Digitalisering av Estlandssvensk ordbok

Forfatter: Lars Törnqvist

Kilde: Nordiske Studiar i Leksikografi 7, 2005, s. 331-338
Rapport frå Konferanse om leksikografi i Norden, Volda 20.-24. maj 2003

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive



Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Lars Törnqvist

Digitalisering av *Estlandssvensk ordbok*

The dictionary of Swedish dialects in Estonia by Nils Tiberg is a manuscript consisting of 88 000 octavo-sized cards. The aim of this project is to make a database, accessible through the Internet, containing search indexes and digital images of the cards.

«Estlandssvensk ordbok» av Nils Tiberg är ett av de allra bäst genomarbetade svenska dialektordboksmanuskripten. Verket redovisar ordförrådet i de svenska dialekter som talades i Estland fram till andra världskriget, med detaljerade uppgifter om uttal, böjningsformer, morfologisk variation och betydelser hos ord och fraser. Kontakten med estniska språket är också belyst i materialet. Ordboksmanuskriptet består av ett kortregister med cirka 88 000 kort hos Språk- och folkminnesinstitutet (SOFI) i Uppsala. För att göra manuskriptet tillgängligt i form av en ordboksdatabas med åtkomst via Internet har ett digitaliseringsprojekt påbörjats. Projektet omfattar scanning av kortregistret, registrering av sökbara data samt uppbyggnad av databasen.

Bakgrund

Under 2001 och 2002 pågick ett projekt med syftet att utarbeta förslag till standardiserad digital lagring av dialekttexter. Projektet, som finansierades av Riksbankens jubileumsfond, leddes av professor emeritus Benny Brodda vid Stockholms universitet i nära samarbete med Språk- och folkminnesinstitutet (SOFI) i Uppsala. Huvuddelen av arbetet ägnades åt metadata för transkriptioner av inspelningar och kodning av text med det svenska landsmålsalfabetet, men inom ramen för projektet studerades också möjligheterna till digital lagring av dialektordböcker och arbetsmaterial till sådana. Tanken var att starta ett pilotprojekt med en lagom stor dialektordbok som skulle göras tillgänglig och maskinellt sökbar via Internet.

I SOFI:s arkiv finns flera ordboksmanuskript i form av kortregister. Som lämpligt material för pilotprojektet föreslog forskningssekreteraren vid SOFI, docent Lars Bleckert, Nils Tibergs ordbok över de estlandssvenska dialekterna. Det fanns flera skäl att välja just den ordboken. Ett skäl är att de estlandssvenska dialekterna är mycket intressanta i allmänt dialektologiskt perspektiv, vilket innebär att många forskare på olika orter i Sverige, Finland och Estland har behov av att komma åt materialet. Ett annat skäl är att ordboksmanuskriptet håller mycket hög kvalitet, vilket gör det lämpligt som studieobjekt för lexikografistuderande och förebild för framtida dialektlexikon.

Estlandssvenskar och estlandssvenska

Ända sedan medeltiden har det funnits en svenskspråkig allmogebefolkning i Estlands kustområden. Det har aldrig varit någon stor folkgrupp; när den var som störst i slutet av 1500-

talet omfattade den cirka 12 000 personer. Vid folkräkningen 1934 fanns 7 641 svensktalande invånare i Estland. De flesta flydde till Sverige under andra världskriget, och numera uppgår antalet estlandssvenskar till några hundra, av vilka många inte använder svenskan som vardagsspråk. Efter kriget har den estlandssvenska kulturen främst levt vidare i Sverige, inte minst genom aktivt föreningsliv.

De estlandssvenska dialekterna uppvisar en del ålderdomliga drag, exempelvis bevarade diftonger och pluralböjning av verb. I fonologiskt hänseende har de åtskilligt gemensamt med finlandssvenska dialekter. Över huvud taget finns likheter i ett eller annat avseende med svenska dialekter runt större delen av Östersjön, även så långt bort som i Kalix längst norrut i Bottenviken. I ett helhetsperspektiv präglas dock estlandssvenskan snarare av isolering från den allmänna språkutvecklingen i Sverige och kontakt med estniska dialekter (på Runö även med lettiska dialekter).

Den svenska bebyggelsen i Estland bestod av flera olika områden, vilka motsvaras av dialektala huvudvarianter (figur 1). Till Estlandssvenskan räknas också den dialekt som talades i Gammalsvenskby i Ukraina, dit flertalet av Dagösvenskarna tvångsförflyttades år 1781.

- Runska eller runösvenska på Runö i Rigabukten.
- Nuckösvenska på Nuckö och i kommunerna Pasklep, Sutlep, Rickul och på Odensholm.
- Ormsösvenska på Ormsö.
- Rågömålen på Stora och Lilla Rågö.
- Vippalasvenska och Korkissvenska i Vippal respektive Korkis.
- Dagösvenska på Dagö.
- Gammalsvenskbymålet i Gammalsvenskby i Ukraina.
- Nargösvenska på Nargö.

Ordboksmanuskriptet

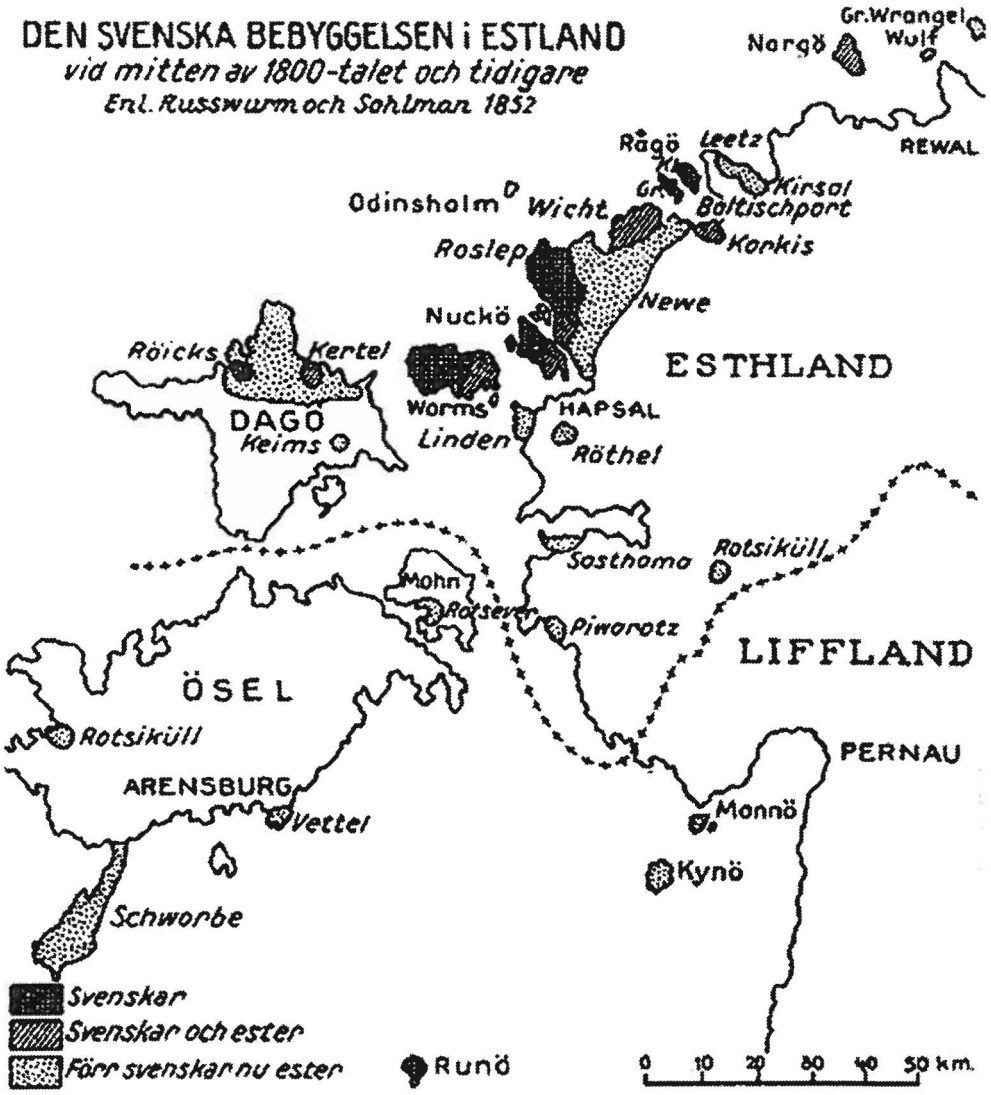
Nils Tibergs estlandssvenska ordbok är baserad på väl dokumenterade språkprov som Tiberg själv eller hans medarbetare har samlat in vid fältstudier, huvudsakligen mellan 1929 och 1945. Tiberg arbetade vidare med bearbetningen av ordboksmaterialet ända till sin död 1980. Därefter har arbetet förts vidare av flera medarbetare, av vilka den viktigaste var Ann-Mari Cronström, som omkom i Estoniakatastrofen 1994.

Ordboksmanuskriptet består av 88 000 kort i oktavformat. Korten har förtryckta ortnamn men är i övrigt skrivna för hand. Uttal anges mycket noggrant med det svenska landsmålsalfabetet.

För varje lemma finns ett huvudkort som upptar uttals- och böjningsformer i de olika estlandssvenska dialekterna (figur 2). Efter huvudkortet kommer sedan ett varierande antal kort av annan typ: betydelsekort för varje betydelse (figur 3), fraskort för fasta uttryck, kommentarkort med upplysningar om exempelvis fonetiska eller grammatiska förhållanden samt skisskort med teckningar eller andra illustrationer.

Projektidé

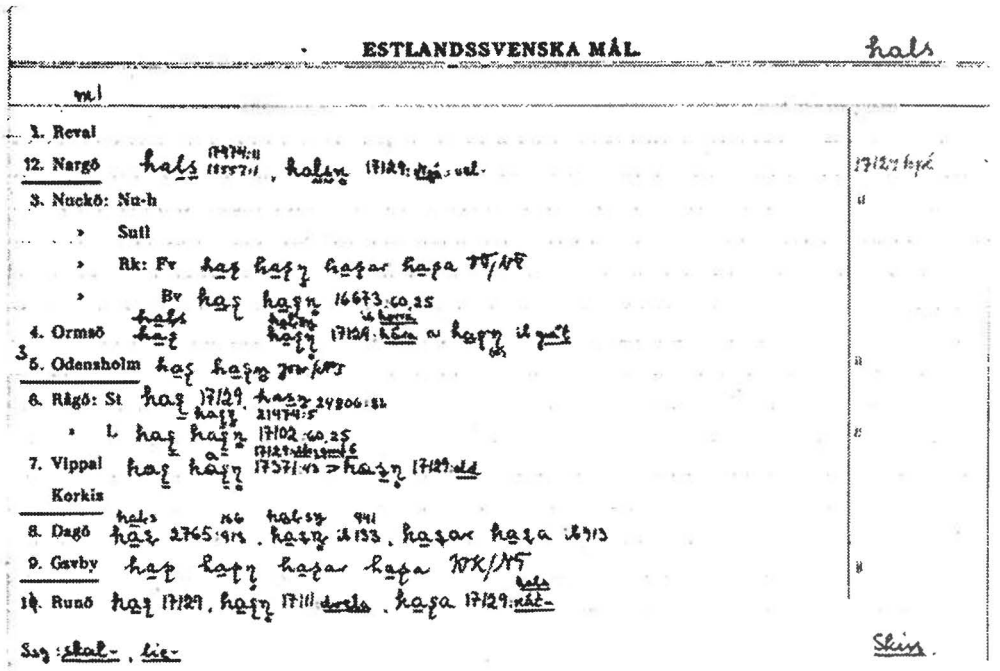
Hur kan man göra detta material åtkomligt för sökning med mycket begränsade ekonomiska



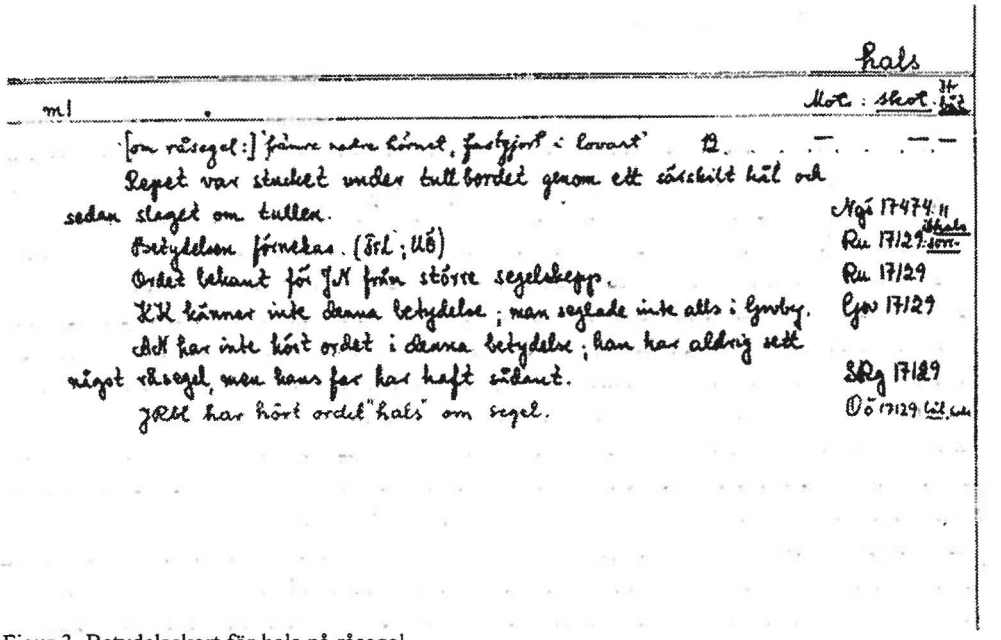
Figur 1. Estlandssvenskans utbredning.

resurser till förfogande? Manuell registrering av texten på korten är mycket tidskrävande och därmed kostsam. OCR-inläsning av texten hade varit ett alternativ om texten hade varit tryckt eller maskinskriven, men dagens OCR-program klarar inte handskrift. Att det delvis är fråga om språkvarianter som starkt avviker från svenskt standardspråk, flera olika handstilar och ett mycket speciellt fonetiskt alfabet gör det hela ännu svårare. Vår lösning blev att scanna korten som bilder och att manuellt registrera metadata för inläggning i en sökdatabas.

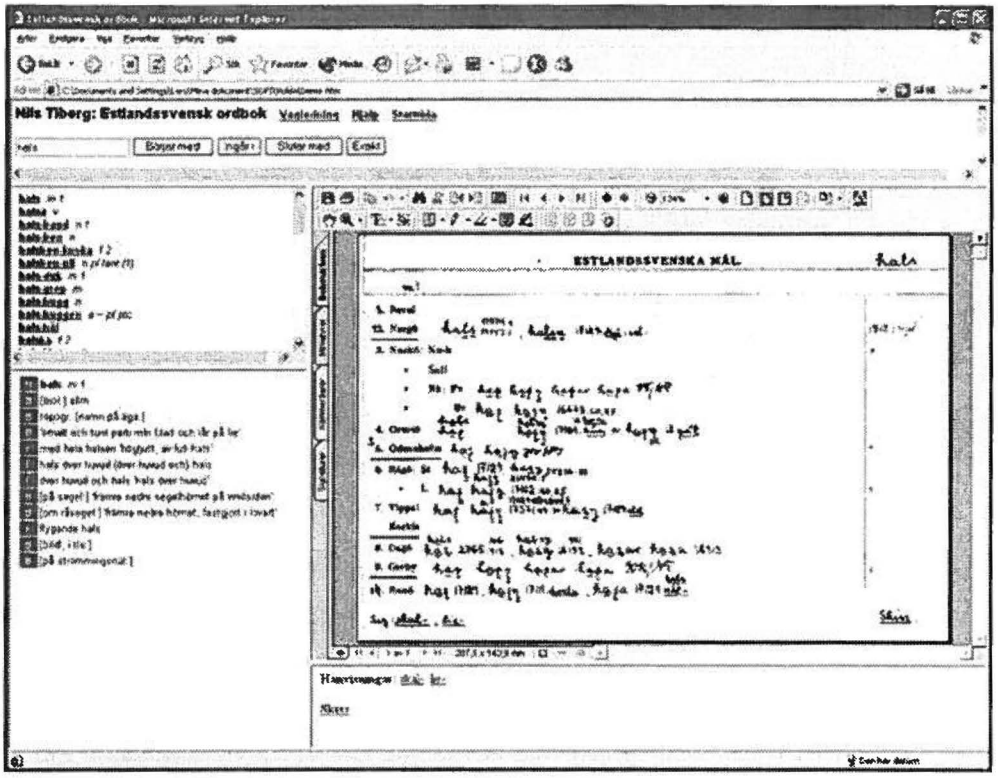
Vi tänker oss att sökgränssnittet skall se ut ungefär som i figur 4. I den översta rutan finns sökfält, knappar för olika söksätt och länkar till hjälpsidor. Användaren skriver in sökordet –



Figur 2. Huvudkort för substantivet *hals*.



Figur 3. Betydelsekort för hals på råsegel.



Figur 4. Sökgränssnitt för ordboken. Den översta breda rutan innehåller sökfält, knappar för olika sökfält samt länkar till hjälptexter. Den övre vänstra rutan innehåller träfflista och den nedre vänstra rutan innehåller översikt över de kort som avser ett lemma. I den övre högra rutan visas ett kort i taget och i den nedre högra rutan finns länkar till andra kort.

eller en del av det – i sökfältet och trycker på knappen för aktuellt söksätt. Om man skriver «hals» i sökfältet och trycker på knappen «Börjar med» får man upp alla ord som börjar med *hals* i träfflistan den övre vänstra rutan: *hals, halsa, hals-band, hals-ben* och så vidare. Med knapparna «Ingår i» och «Slutar med» kan man enkelt söka på mellanled respektive slutled. Trycker man på knappen «Exakt» får man träff på bara den teckensträng man har skrivit i sökfältet, alltså *hals*.

I träfflistan klickar man på det lemma man vill studera. Då kommer det upp i den nedre vänstra rutan en lista på alla kort som avser lemmat. Listan motsvarar en översikt över mikrostrukturen i en vanlig ordboksartikel. Den är uppställd i samma ordning som de fysiska korten är insorterade i sina lådor. Först kommer huvudkortet. Därefter kommer de olika betydelseavdelningarna, var och en bestående av ett betydelsekort och ett varierande antal fraskort, kommentarkort och skisskort. I listan anges korttyp och uppgifter om kortets innehåll, till exempel fras i rikspråklig form («över huvud och hals»), användningsområde («på segel») eller betydelseangivelse («högljutt, av full hals»).

När man klickar på en rad i listan får man upp en bild på det aktuella kortet i den stora

rutan till höger. Genom att bilderna är lagrade i PDF-format kan man zooma in och ut efter behov, så att man kan se hela kortet för att få överblick eller göra en delförstoring av ett par ord för att kunna tolka otydliga noteringar bättre.

Många kort innehåller hänvisningar till andra kort. Dessa kort nås enkelt genom att man klickar på hyperlänkarna i den nedersta högra rutan.

Dataregistrering

Den ordboksdatabas som beskrivs ovan är en vision. För att visionen skall förverkligas krävs ett omfattande arbete. Dels måste alla kort lagras som digitala bilder, dels måste metadata för alla korten registreras.

För att man skall få kontroll över hela materialet måste varje kort ha en unik identitet. Denna består av ett löpnummer som stämplas på baksidan av kortet. Samma nummer anges i de poster som innehåller metadata för respektive kort. Numret används sedan internt i databassystemet för att koppla ihop posten i databasen, träfflistan, kortlistan, bildfilen och hyperlänkarna.

Korten scannas i nummerordning. Varje kort lagras som en bildfil i TIFF-format, vars namn består av ett prefix och kortets nummer. Experiment med olika upplösningar visade att det krävdes minst 200 dpi för att ge tillfredsställande återgivning av de minsta detaljerna i handskriften. Högre upplösning medförde markant större filstorlek utan motsvarande förbättring av läsligheten. Återgivning av gråtoner förbättrade inte läsligheten. Tvärtom gav svartvita bilder med en pixels färgdjup (alltså med bara svart och vitt) skarpare och mera lättläst text, förutsatt att gränsvärdet mellan svart och vitt var rätt inställt vid scanningen. Vissa kort innehåller skrift i olika färger. Bildfiler med färginformation är dock flera gånger så stora som svartvita bildfiler. Färgerna är som regel inte nödvändiga för tolkningen av texten, och det bedömdes därför att färgbilder av korten inte var nödvändiga. Slutsatsen blev allstå att korten skulle scannas i svartvitt med en upplösning av 200 dpi. De TIFF-bilder som scannern lämnar ifrån sig konverteras efteråt till PDF-format.

Metadata för korten skall lagras i XML-format, vilket gör det enkelt att bygga upp en databas av materialet. XML-formatet är dock ganska otympligt som registreringsformat, så därför gjordes ett förenklat format för dataregistreringen. Texten skrivs in i en mall där varje fält föregås av en tagg (figur 5). Taggarna har följande betydelser:

KORT = kortets nummer
HGNR = homografnummer
UORD = uppslagsord
MMNR = momentnummer (för betydelsemoment)
GRAM = grammatiska uppgifter (ordklass, genus, deklination)
KTYP = korttyp
  H = huvudkort
  B = betydelsekort
  F = fraskort
  K = kommentarkort
S = skisskort
FRAS = fras

KORT 572
KTYP B
HGNR
UORD hals
MMNR
GRAM m
FRAS
BETY 'smalt och tunt parti mln blad och «lår» på «lie»'
REFE Syn: lie-hals
=
KORT 573
KTYP F
HGNR
UORD hals
MMNR
GRAM m 1, i fras:
FRA Smed hela halsen
BETY 'högljutt, av full hals'
REFE Jfr: ljud; Jfr: hårt
=
KORT 574
KTYP F
HGNR
UORD hals
MMNR
GRAM m 1, i fras:
FRAS hals över huvud
BETY
REFE Syn: (över huvud och) hals; Syn: ars (över huvudet); Jfr: plötsligt
=

Figur 5. Utdrag ur registreringsfil för metadata.

```
<kort>
<kortnummer>4711</kortnummer>
<korttyp>F<korttyp>
</kort>
<lemma>
<homografnummer>II<homografnummer>
<uppslagsord>hälsa</uppslagsord>
<momentnummer>2</momentnummer>
<grammatik>v</grammatik>
</lemma>
<fraser>
<fras>hälsa hem</fras>
</fraser>
<betydelse>vara illa ute</betydelse>
<hänvisningar>
<hänvisning>
<hänvisningstyp>Jfr</hänvisningstyp>
<kortnummer>8898</kortnummer>
<lemma>
<homografnummer>III</homografnummer>
<uppslagsord>råka</uppslagsord>
<momentnummer>3</momentnummer>
<grammatik>v</grammatik>
</lemma>
</hänvisning>
</hänvisningar>
```

Figur 6. Utdrag ur XML-fil med metadata.

BETY = betydelsebeskrivning
REFE = referensfält för jämförelser, synonymer, motsatsord, sammansättningar och liknande.

För textregistreringen används ett vanligt ordbehandlingsprogram. När texten är färdigregistrerad konverteras den till XML-format med ett enkelt konverteringsprogram (figur 6).

Nuläge och framtid

Scanning och textregistrering har utförts vid SOFI i Uppsala. Större delen av arbetet har gjorts av en lönebidragsanställd person, vars anställdning upphörde under våren 2003. Då hade alla kort för ord med för begynnelsebokstaven H scannats och metadata för dessa kort har registrerats. Korrekturläsning av texten återstår.

Vad som händer med projektet i framtiden beror på möjligheterna att finansiera fortsatt arbete. De interna resurserna inom SOFI är helt otillräckliga, varför extern finansiering i någon form är nödvändig.

Projektet kring Tibergs ordbok är ett pilotprojekt. Detta innebär att idéer och erfarenheter från projektet skall läggas till grund för kommande projekt av liknande slag. I arkivet hos SOFI finns flera ordboksmanuskript av liknande typ för olika dialektområden i Sverige. Dessutom finns det omfattande materialet till Ordbok över Sveriges dialekter, bestående av mer än åtta miljoner kort, vilka skulle kunna göras tillgängliga för forskare via Internet på ungefär samma sätt som skisserats ovan. Detta skulle kunna ge avsevärt bättre möjligheter för forskare att komma åt upptecknat material som rör svenska dialekter, vilket skulle innebära ett stort lyft för svensk språkforskning.