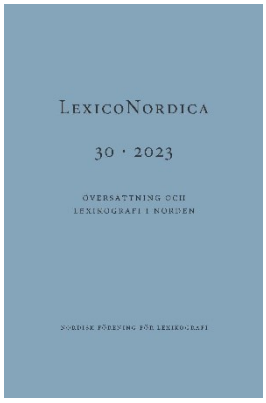
LexicoNordica

Titel:	Lexikografiska resursers betydelse i utvecklingen av språkteknologiska verktyg för minoritetsspråk
Forfatter:	Marie Mattson & Magnus Ahltorp
Kilde:	LexicoNordica 30, 2023, s. 75-94
URL:	https://tidsskrift.dk/lexn/issue/archive



Lexikografiska resursers betydelse i utvecklingen av språkteknologiska verktyg för minoritetsspråk

Marie Mattson & Magnus Ahltorp

The Swedish Language Act states that the public sector is responsible for protecting the Swedish national minority languages. Since these languages typically have few speakers, commercial actors do not create the language technology necessary in an increasingly digitalised society. One way for the public sector to facilitate creation of such technology is by making lexicographic resources available to the public. In this article, we present a survey of existing resources and guidelines for new resources, such as the early inclusion of a computational linguistic perspective.

1. Introduktion

Sedan 2009 har Sverige en språklag som bland annat definierar fem nationella minoritetsspråk: finska, jiddisch, romani chib, meänkieli och samiska. Enligt lagen har ”[d]et allmänna […] ett särskilt ansvar för att skydda och främja de nationella minoritetsspråken” (Språklagen).

Att det allmänna har ett särskilt ansvar för att skydda och främja de nationella minoritetsspråken kan betyda många saker, som exempelvis att höja språkens status bland ungdomar, skriva och publicera fler texter på språket ifråga och normera skriftspråket med hjälp av skrivregler. Rapporten *Språklagen i praktiken* (Institutet för språk och folkminnen 2011) lyfter bland annat översättning av texter till de nationella minoritetsspråken som ett sätt att stärka dem. Inte bara för att öka tillgängligheten av texternas innehåll för talare av de nationella minoritetsspråken, utan även för att det synliggör de nationella minoritetsspråken i flera sammanhang.

Trots att statliga myndigheter och kommuner förväntas över-

sätta sina texter till de nationella minoritetsspråken, är tillgången till språkliga resurser och verktyg begränsad. Det kan leda till sämre och dyrare översättningar, särskilt för de minoritetsspråk som inte har helt normerade skriftspråk och där det kan vara svårt att hitta auktoriserade översättare.

I vårt alltmer digitaliserade samhälle glöms den digitala aspekten ofta bort när det handlar om att främja minoritetsspråken. Idag sker vår kommunikation i stor utsträckning med digitala hjälpmedel vilket gör det ännu viktigare att också dessa språk får tillgång till digitala verktyg. Obefintliga eller undermåliga digitala språkverktyg kan leda till att talare av språken avstår från att använda dem. Bristen på digitala verktyg och resurser för de nationella minoritetsspråken utgör därför ett stort hot mot språkens överlevnad (Borin et al. 2012).

För att främja den digitala utvecklingen av de nationella minoritetsspråken behövs inte bara mer teknik, utan även tillgång till språkresurser i digital form som kan ligga till grund för utveckling av språkteknologi för det aktuella språket. Exempel på språkresurser är texter på språket i fråga i skriven eller talad form och lexikografiska resurser som ordböcker och termlistor. För flera av de nationella minoritetsspråken är behovet av flera, större och bättre lexikografiska resurser stort. Texter och lexikografiska resurser är ett hjälpmedel både vid språknormering och översättning samt ett underlag för utveckling av språkteknologiska verktyg.

De språkteknologiska verktygen, i sin tur, är också viktiga redskap vid språknormeringen och en viktig resurs för översättare. Några exempel på språkteknologiska verktyg är tangentbord för dator och mobil, program för stavningskontroll, grammatikkontroll och maskinöversättning samt talteknologi (talsyntes och taligenkänning).

Eftersom de nationella minoritetsspråken är små och majoriteten av deras talare även talar ett majoritetsspråk, är det kommersiella intresset för att utveckla lexikografiska resurser och språktek-

nologiska verktyg svalt. Därför har det allmänna, genom Sveriges språkmyndighet Institutet för språk och folkminnen (Isof), ett ansvar att främja den digitala utvecklingen genom att tillhandahålla digitala resurser för de nationella minoritetsspråken.

Den här artikeln går igenom varför lexikografiska resurser är viktiga för utvecklingen av språkteknologiska verktyg, med fokus på de nationella minoritetsspråken finska, meänkieli, romani chib (romska) och jiddisch[1]. Utifrån det diskuteras hur en bra lexikografisk resurs bör se ut för att vara användbar för utveckling av språkteknologiska verktyg.

2. Bakgrund

2.1 Hotade språk och språkteknologi

Det finns ett flertal sätt att mäta språks livskraft. Samtliga språk som diskuteras i den här artikeln, förutom finska, räknas som hotade eller sårbara enligt UNESCO:s *World Atlas of Languages* (UNESCO 2021).

Ett annat sätt är att använda sig av termen LOL, vilket står för ”Literate, Official, and with Lots of users” (Dahl 2015), där ”Literate” betyder att det finns en livskraftig produktion av t.ex. tidningsartiklar och böcker. Dahl (2015) använder Wikipedia för att mäta vilka språk som uppfyller ”Literate”-kriteriet, officiell status i minst ett land för att uppfylla ”Official”-kriteriet och minst en miljon talare för ”Lots of users”. När Dahl kombinerar alla tre kriterier är det endast 57 av alla världens språk som uppfyller samtliga kriterier. Om dessa kriterier inte uppfylls blir det svårt för ett språk att överleva i dagens skriftspråkstunga samhälle utan någon form av stöd (Dahl 2019). Bland de fem svenska nationella minoritets-

1 Samiska språk diskuteras inte i artikeln eftersom de samiska språken inte ingår i Isofs uppdrag.

språken uppfyller endast finska LOL-kriterierna (vilket dock inte omfattar de sverigefinska varieteterna). Romani chib har i och för sig många talare, men når inte upp till Dahls (2015) ”Literate”-kriterium.

Inom forskning och den kommersiella marknaden tenderar det att uppstå en LOL-bias. Eftersom flera av de nationella minoritetsspråken inte har ”Lots of users” saknas generellt intresse från kommersiella aktörer för att ta fram språkteknologiska verktyg och resurser. Även när ett språk har många talare krävs oftast även en ekonomisk och/eller politisk pondus för att ett intresse ska uppstå. Om textproduktionen är mindre omfattande för ett språk (”Literate”-kriteriet), blir det svårare att både beforska språken och utveckla språkteknologi. Arbetet försvåras ytterligare av att skriftspråken tenderar att vara relativt unga och inte alltid helt normerade.

2.2 Språkteknologi i en alltmer digital värld

I takt med att samhället blir mer digitalt, blir även språket mer digitalt. Sådant som tidigare förmedlades via direkt tal kan nu skrivas på mejl, sms eller läsas in som röstmeddelande. Tidigare har skriftspråk använts i färre och ofta mer formella sammanhang: i litteratur, på arbetsplatsen och i tidningar. Självklart är informell skrift inte ett nytt fenomen, men det är mycket vanligare och mer utbrett idag än tidigare.

Digitaliseringen får stora konsekvenser, både positiva och negativa. Den kan bidra till att göra språket mer tillgängligt, till exempel finns det idag bättre möjligheter för människor med läs- och skrivsvårigheter att enklare ta till sig viktig samhällsinformation i skriven form. Den ger också ytterligare möjligheter till andraspråksinlärning och till att tillgodogöra sig information på andra språk, exempelvis med hjälp av automatisk textning, digitala ordböcker och digitala lärplattformar. Men det ökade an-

vändandet och behovet av avancerad språkteknologi leder också till att ojämlikheterna mellan majoritetsspråk och minoritetsspråk ökar.

För att skriftspråket ska kunna användas måste de tekniska förutsättningarna finnas. I en värld där de tekniska hjälpmedel som används är penna och papper ställer ett språk oftast inte några speciella krav på tekniken. Skrivmaskinen, och tangentbordet som kom med den, uteslöt däremot språk. Även språk med någorlunda lika skriftsystem kan vara svåra att skriva på en skrivmaskin av fel typ. Till exempel måste den som skriver svenska på en skrivmaskin med engelskt tangentbord manuellt sätta dit prickar och ringar över *å*, *ä* och *ö* eller helt låta bli att använda dessa tecken.

Datorer och mobiltelefoner var i början mycket svåra att använda om hårdvara och mjukvara inte köptes för rätt språk. Här gick det inte ens att rita dit ringar över *å*, utan andra tecken fick användas.

Idag är mobiltelefoner ofta konstruerade för många språk, så att exempelvis en telefon som köpts i Italien kan skriva svenska tecken, bara den ställs om till svenskt tangentbord. Detta betyder inte bara att telefonen visar rätt tecken och tillåter inmatning av dessa, utan även att tangentbordet använder en svensk språkmodell. Språkmodellen innehåller statistik om vilka tecken som ofta förekommer tillsammans och kan därför förutsäga vilket tecken som är det mest troliga givet användarens oprecisa inmatning, något som är till stor hjälp på små skärmar. Den hjälper även till med stavning och gissar vilka ord som kommer härnäst för att underlätta skrivandet (Fowler et al. 2015).

Men vad händer när denna mjukvara inte finns för ett språk? För att ta ett konkret exempel kan meänkieli skrivas på ett finskt eller svenskt tangentbord, tecknen finns där, men språkmodellen är fel. Telefonen kommer då att försöka tvinga på användaren finsk respektive svensk stavning.

Nu när alltmer av vår kommunikation sker via digitala medel

riskerar därför små språk att användas i färre sammanhang och därmed bli ännu mer hotade.

Visst kan språk användas och överleva utan tillgång till språkteknologi, men idag sker såpass mycket av vår kommunikation digitalt och i skrift att det är en förutsättning för att språken ska överleva och frodas under en längre tid.

Talare av ett minoritetsspråk behärskar i allmänhet även landets eller regionens majoritetsspråk, som oftare går att använda digitalt utan hinder. Det är därför troligt att talare av minoritetsspråk istället använder majoritetsspråket i många sammanhang eftersom den tekniken är mer lättillgänglig och bättre. Det är förstås inte ett problem i sig att minoritetsspråkstalare även talar och skriver på majoritetsspråket i stor utsträckning, men avsaknad av språkteknologiska verktyg får inte vara ett hinder för den som vill eller behöver använda sitt minoritetsspråk. För att det ska vara möjligt måste språkteknologin utvecklas.

2.3 Sveriges nationella minoritetsspråk och tillgång till språkteknologi

De nationella minoritetsspråken i Sverige riskerar att lida av domänförluster, färre talare och i värsta fall kan de även dö ut om de inte kan användas digitalt (Borin et al. 2012). De har dock olika god tillgång till språkteknologiska verktyg. Inom projektet European Language Equality (ELE) kartlades situationen för de nordiska minoritetsspråken (Nørstebø Moshagen et al. 2022).

Finskans situation skiljer sig något från de andra språkens eftersom den även är majoritetsspråket i Finland. Där finns cirka fem miljoner talare och där har en god språkteknologisk infrastruktur redan byggts upp. Efter finska är jiddisch det språk som har kommit längst av dessa fyra språk. Grundläggande språkteknologiska verktyg som tangentbord och stavningskontroll finns i alla fall till viss del för jiddisch. Bland de mer avancerade språktek-

nologiska verktygen finns tillgång till maskinöversättning mellan engelska och jiddisch.

Vad gäller såväl meänkieli som romska pågår ett arbete med att ta fram språkmodeller som både kan användas som stavningskontroll och annoteringsverktyg för korpusar. Båda språken saknar tangentbord anpassade till respektive språk och den som skriver på meänkieli eller romska behöver därför använda ett tangentbord som är avsett för ett annat språk.

Förutom brist på språkteknologiska verktyg och intresse från stora företag är flera av de nationella minoritetsspråken mycket mindre beforskade än stora världsspråk som engelska, eller de nordiska majoritetsspråken. Mycket av det beror förstås på att engelska har många talare, men även att det finns enorma mängder material att forska på. Material på engelska finns ofta lättillgängligt, omfattar väldigt många olika domäner och är väldigt stort mätt i antalet ord.

De digitala språkresurser som finns för de nationella minoritetsspråken är i regel begränsade, kommer från en eller ett par upphovspersoner och täcker sällan fler än en handfull domäner.

3. Lexikografiska resurser och andra språkresurser

De resurser som finns kommer från olika håll och har därför olika bakgrund och förutsättningar vad gäller digitalisering och i vilken mån de kan tillgängliggöras som öppna data. Öppna data handlar dels om att data är gratis tillgängligt för alla, dels om hur det är tillåtet att vidareutnyttja data (licensen). Isof strävar efter att de språkresurser som publiceras ska ha en så öppen licens som möjligt. Den mest öppna licensen är CC0 (data helt fritt från begränsningar vid vidareutnyttjande) som används för flera resurser. Även CC-BY (krav på att skaparen nämns vid vidareutnyttjande) är en vanlig och öppen licens som används vid Isof.

Öppna språkresurser publiceras på sidan Isofs digitala språkresurser (härefter IDS) medan metadata för språkresurserna publiceras på Svensk nationell datatjänst.

3.1 Vad är digitala språkresurser?

Digitala språkresurser för ett visst språk kan vara vad som helst som är producerat på språket i fråga och finns tillgängligt i digital form, exempelvis lexikografiska resurser (ordlistor, ordböcker och termlistor). Andra exempel på språkresurser är texter (inklusive översatta texter) och ljudinspelningar (gärna med tillhörande transkription). Digitala språkresurser kan bland annat användas för att:

1. utveckla språkteknologiska verktyg med hjälp av maskininlärning,
2. utveckla språkteknologiska verktyg med andra metoder, till exempel regelbaserade metoder,
3. normera skriftspråket,
4. bygga korpusar som kan användas inom forskning och språkvård.

Lexikografiska resurser är helt nödvändiga för en rad språkteknologiska verktyg, men beroende på vilken typ av verktyg som ska tas fram ställs det olika krav på den lexikografiska resursen. Till exempel är den enklaste språkmodellen för ett mobiltangentbord en lista med hur troligt ett tecken är givet det tecken som kommer innan. För att förbättra den något krävs bara en enkel ordlista med lemman, utan morfologiska uppgifter. Innehåller ordlistan däremot också morfologisk information och ordklass blir den mycket mer användbar.

När det gäller stavningskontroll är den enklaste formen av ordlista, utan morfologisk information, i princip helt oanvändbar

för de flesta språk, eftersom stavningskontrollen då antingen kommer ge många falska positiver (om de morfologiska reglerna är konservativa) eller många falska negativer (om de morfologiska reglerna genererar alla möjliga former baserat på lemmaformen). För att ta ett exempel på svenska ska *kakor* godkännas för att substantivet *kaka* finns som lemma, men *bakor* ska inte godkännas trots att verbet *baka* finns, eftersom den formen inte finns för verbet.

Exempelmeningar i lexikografiska resurser kan också ge värdefull information, eftersom de kan ses som en minikorpus.

3.2 Lexikografiska resurser och språkteknologi som finns för de nationella minoritetsspråken idag

Isof arbetar aktivt för att ta fram och tillgängliggöra lexikografiska resurser på de nationella minoritetsspråken. De flesta lexikografiska resurser är i första hand anpassade för att vara användbara för människor. För att resurserna ska vara användbara för datorer och språkteknologisk utveckling görs en språkteknologisk anpassning av datat.

3.2.1 Jiddisch

För jiddisch finns en del digitala verktyg och resurser. Språket talas i flera länder och som framgår i avsnitt 2.3, finns maskinöversättning mellan engelska och jiddisch. Däremot saknas motsvarande verktyg för svenska–jiddisch. Det finns tangentbord och stavningskontroll för jiddisch, dock inte alltid anpassade efter svenska förhållanden. Även textanalysverktyg och talteknologi saknas (Nørstebø Moshagen et al. 2022:39).

Bland de språkliga resurserna finns enspråkiga och flerspråkiga korpusar samt *Jiddisch–svensk–jiddisch ordbok*. Ordboken publicerades första gången 2005, som nyutgåva 2020 (Kerbel et al. 2020) och digitaliserades av Isof 2022. Ordboken har tillgänglig-

gjorts som *Digital jiddisch–svensk–jiddisch ordbok* och finns idag både online och offline (Ahltorp et al. 2022).

Förutom ett webbgränssnitt publicerades även rådatat på IDS med licensen CC0. Filen består av ett relativt simpelt JSON-format som innehåller ord på svenska, ordets motsvarighet på jiddisch med hebreiska tecken och ordet på jiddisch translittererat med latinska tecken. Dessutom är orden ordklasstaggade. Se figur 1 för ett exempel på rådatat. I datat står ”sv” för ”svenska och ”yi” för ”jiddisch”. Taggen ”graminfo” står för ”grammatisk information” och anger ordklass.

```
{
  "ID": "48890",
  "sv": {
    "ord": "omgärda",
    "graminfo": "vb"
  },
  "yi": {
    "ord": {
      "Hebr": " [-געריגלט] ארומרינגלען",
      "Latn": "arúmringlen (-geringlt)"
    }
  }
}
```

Figur 1. Utdrag ur rådata för *Digital jiddisch–svensk–jiddisch ordbok*.

3.2.2 Meänkieli

För meänkieli saknas grundläggande resurser som tangentbord och stavningskontroll (Nørstebø Moshagen et al. 2022:24). Som tidigare nämnts kan talare av meänkieli använda ett svenskt eller finskt tangentbord, men får då stavningskontroll på svenska respektive finska.

Bland Isofs lexikografiska resurser på meänkieli finns *Meänkieli–svensk–meänkieli ordbok* som bland andra STR-T (Svenska Tornedalingars Riksförbund – Tornionlaaksolaiset), Meän akateemi och Isof har tagit fram. Den är finansierad av Isof och Statens Kulturråd. Ordboken är digital och kan användas i ett sökgränssnitt. Förutom användargränssnittet finns ordboken tillgänglig som öppna data, precis som jiddischordboken, men i ett XML-format.

Orden i ordboken är märkta med ordklass (”pos”) och i de fall det är relevant är de även märkta med en eller flera geografiska varieteter: Gällivare (Je), Kiruna (Kie) och Tornedalen (To). I figur 2 finns ett exempel från en av filerna med meänkieli–svenska för ordet *delad*. Notera att rådatat ser mycket annorlunda ut i förhållande till jiddischordboken.

```
<e>
   <lg>
      <l pos="a">delad</l>
   </lg>
   <mg>
      <tg xml:lang="fit">
         <t geo="To" pos="a">halastu</t>
         <t geo="Kie" pos="a">haljastu</t>
         <t geo="Je" pos="a">halkastettu</t>
      </tg>
   </mg>
</e>
```

Figur 2. Utdrag ur rådata för *Meänkieli–svensk–meänkieli-ordbok*.

3.2.3 Finska

Eftersom finska inte bara är ett minoritetsspråk i Sverige, utan även ett majoritetsspråk i Finland, råder det ingen brist på finska och finsk–svenska språkteknologiska verktyg. Trots detta finns förstås behov av ordlistor med ord som är specifika för det svenska samhället.

Det finns 23 lexikografiska resurser för finska på IDS. Resurserna kan delas in i två kategorier: *Sverigefinska ordlistor* och *Lexin*.

De sverigefinska ordlistorna är 22 till antalet, är olika stora och samtliga mellan sverigesvenska och sverigefinska. Den minsta listan, Muminfigurer-listan, består av 26 uppslag. Den svensk–finska socialordlistan är den största och innehåller cirka 3 700 uppslagsord. Även den svensk–finska skolordlistan är stor och består av 1 836 uppslagsord. De flesta ordlistorna är något mindre till storleken, mellan 100 och 300 ord. Majoriteten av ordlistorna kommer från den finska tidskriften *Kieliviesti* som Isof publicerar två gånger per år (fram till 2020 publicerades fyra nummer per år).

En av ordlistorna, den svensk–finska omsorgsordlistan, har ett väldigt simpelt JSON-format, den enda ordlistan med det formatet. Ordlistan har endast märkningen ”sv” för de svenska orden, och ”fi” för de finska motsvarigheterna. Övriga ordlistor har ett XML-format som i grunden är samma som *Lexins* (se strax nedan).

Den sista lexikografiska resursen som finns för finska hos Isof är *Lexin* – Lexikon för invandrare. *Lexin* finns som öppna data på finska och 19 andra språk. *Lexin* använder ett XML-format, som utöver uppslagsord och översättning (som finns i de andra svensk–finska lexikografiska resurserna), även innehåller betydelseförklaringar, kommentarer, svenskt uttal och grammatisk information. Den grammatiska informationen består av ordklass och böjningar av det svenska ordet. *Lexin* innehåller dock inga översättningar till sverigefinska, utan enbart finlandsfinska.

De sverigefinska ordlistor som använder lexinformatet har samma struktur som *Lexin* men mycket mindre information. De innehåller bara uppslagsordet på svenska och dess motsvarighet på sverigefinska. Mängden information är därför samma som för omsorgsordlistan, endast formatet skiljer.

3.2.4 Romska

För romska saknas grundläggande språkteknologiska verktyg, bland annat tangentbord. I Sverige talas flera dialekter av romska som ofta har separata lexikografiska resurser. Isof har lexikografiska resurser för följande dialekter:

- polsk romska (1 resurs),
- resanderomska (1 resurs),
- arli (6 resurser),
- kalo (7 resurser),
- kelderasch (5 resurser),
- lovari (5 resurser).

Totalt finns det åtta olika ordlistor men endast en av dem, omsorgsordlistan, finns på alla romska dialekter. Den finns dock endast i tryckt format.

De ordlistor som finns tillgängliga online finns dels i form av tabeller på Isofs vanliga webbsidor (*Svensk-romska ordlistor*) och dels som öppna data med ett väldigt enkelt JSON-format. Formatet innehåller ingen information utöver det svenska ordet, dess motsvarighet på romska och vilken romsk dialekt som avses. Formatet är snarlikt den svensk–finska omsorgsordlistan.

Den romska coronaordlistan skiljer sig något från de andra eftersom den innehåller tre olika romska dialekter (arli, kelderasch och lovari) i samma fil. Alla andra romska ordlistor har en separat fil för varje dialekt i de fall det finns flera.

Utöver dessa ordlistor finns även *Myndighetstermlistan* som togs fram i Isofs projekt Flersamterm (Institutet för språk och folkminnen 2022). Ett av resultaten från projektet var en flerspråkig termlista, *Flerspråkig basterminologi* (2022), som översattes till bland annat arli och kelderasch. Termlistan finns tillgänglig som öppna data med licensen CC-BY. *Myndighetstermlistan* finns på flera språk utöver svenska och romani (arabiska, finska och eng-

elska). Termlistan använder TBX-formatet, ett XML-format som ofta används i terminologisammanhang och därför är lätt att importera till översättningsprogram. Termlistan innehåller förklaringar och definitioner på svenska.

3.3 Hur ser en bra lexikografisk resurs ut?

Hur ska egentligen en lexikografisk resurs se ut för att vara användbar för språkteknologisk utveckling? Ofta, framförallt vad gäller språk som inte har så mycket resurser, får man nöja sig med det som finns tillgängligt. I de flesta fall är lexikografiska resurser framtagna just med en mänsklig användare i åtanke. Det är förstås naturligt utifrån en lexikografs eller språkvårdares perspektiv. Därför kan det vara bra att redan från början ha med det språkteknologiska perspektivet och i bästa fall samarbeta med en språkteknolog. Även om det inte är möjligt att anlita en språkteknolog finns det några saker som kan vara bra att tänka på när man tar fram nya lexikografiska resurser eller vidareutvecklar befintliga.

Ju mer information som finns i resursen, desto mer användbar kan den vara. Att till exempel ta med uppgifter om ordklass och böjningsformer eller geografisk information kan vara användbart för språkteknologer, språkvårdare och forskare. Även den enklaste typen av lexikografisk resurs – en enspråkig ordlista – kan förstås ha sina användningsområden. Exempelvis kan den ligga till grund för en enkel stavningskontroll, men resursen blir mer användbar om den även innehåller information om ordens böjning.

Det är också viktigt att informationen är strukturerad. I exempelvis Isofs jiddischordbok finns information om genus, böjning och vilket hjälpverb som ska användas tillsammans med uppslagsordet. Däremot är det svårt att använda denna information eftersom den inte framgår explicit i dataformatet, utan är en del av ordet. En utökning av formatet där till exempel genus flyttas till ett separat attribut skulle därför underlätta för vidare användning,

inte bara för den som är intresserad av genus utan även för den som vill ha ordet utan bestämd artikel.

Information om olika varieteter kan också vara användbar för språk som meänkieli och romska som har olika dialekter. Grammatisk information, som till exempel vilken ordklass orden tillhör, kan vara användbart för automatiska annoteringsverktyg. Om det dessutom finns ljudinspelningar som motsvarar orden kan dessa vara användbara för utveckling av exempelvis talsyntes eller taligenkänning. Inspelningar av enstaka ord räcker normalt inte för utveckling av tillfredställande talteknologi, utan en viktig resurs i det sammanhanget är inspelade konversationer (ELE Consortium 2022:31). Däremot är det i praktiken svårt att få täckning av en stor del av ordförrådet på detta sätt. Därför är all form av ljuddata för små språk värdefull.

Det är mycket kostsamt att i efterhand matcha en ljudinspelning med rätt ord eller lägga till information om ordklass och böjning. Därför är det viktigt att planera arbetet så att kostnaden blir låg. Exempelvis är det oftast lämpligt att grammatisk information matas in i samband med att övrigt data matas in för ett ord och att eventuella ljudinspelningar kopplas ihop med ordet redan vid inspelningen.

Utöver det språkliga finns en del tekniska aspekter som är viktiga att tänka på. Självklart måste lexikografiska resurser vara digitalt läsbara och i ett lämpligt format för att vara användbara i ett språkteknologiskt sammanhang. Som tidigare redovisats använder Isof många olika digitala format som var för sig kanske inte ställer till med problem, men antalet format kan i sig vara ett problem. De format som nämns ovan är antingen XML- eller JSON-baserade, men det är viktigt att förstå att XML och JSON bara specificerar den övergripande strukturen. Detaljerna varierar mellan olika XML- och JSON-format, och det finns inget allmänt accepterat format för lexikografiska resurser. Även för terminologiska resurser, där det XML-baserade TBX-formatet är vanligt,

kan TBX upplevas som antingen för komplicerat för en lättare lista, eller inte tillräckligt uttrycksfullt för mer komplexa resurser.

Om ett eget XML-format behöver skapas för resursen är det viktigt att det inte finns några tvetydigheter kring vad de olika taggarna i filen betyder. Det innebär att en tagg i till exempel en XML-fil endast bör stå för en sak konsekvent genom hela filen. Det kan låta självklart, men det är lätt hänt att det glöms bort om man inte har det i åtanke från början.

Ytterligare en aspekt som är viktig att tänka på från början är hur öppen för vidareutnyttjande resursen kan göras. Finns det några upphovsrättsliga skäl till att en resurs behöver en mer begränsad licens eller kan den ha CC0 eller CC-BY som är väldigt öppna licenser? Om man tar upp den frågan med eventuella samarbetspartners redan från början underlättar det distributionen av resursen.

4. Avslutning

Att utveckla goda digitala resurser för de nationella minoritetsspråken innebär lika mycket jobb, eller mer jobb, som för stora majoritetsspråk. Ett av de största hindren är dels att kommersiella aktörer har ett svagt intresse att arbeta med mindre språk, dels att mängden språkdata är betydligt mindre för små språk. Idag pågår ett arbete med att utveckla digitala resurser och verktyg för de nationella minoritetsspråken, bland annat på Giellatekno vid Universitetet i Tromsø och på Isof.

De nationella minoritetsspråken i Sverige har kommit olika långt med språkteknologi och tillgång till digitala lexikografiska resurser, men gemensamt för alla är att det behövs fler och bättre resurser som täcker fler språkliga domäner. Idag finns det lexikografiska resurser med olika format och storlek för alla de nationella minoritetsspråken. På finska finns det relativt gott om

lexikografiska resurser, som *Lexin* och flera stora domänspecifika ordlistor, men på exempelvis romska finns endast kortare domänspecifika ordlistor. På Isof arbetar man inte bara med att ta fram nya lexikografiska resurser, utan även med att digitalisera de som redan finns. Det innebär att antalet lexikografiska resurser som finns tillgängliga som öppna data förmodligen kommer öka något i framtiden. Men för att det ska kunna ske en verklig förbättring behövs fler resurser och översättningar till de nationella minoritetsspråken. De lexikografiska resurserna hos Isof är också mycket heterogena, vilket kan försvåra användningen.

Ur ett översättarperspektiv vore det bra om de flerspråkiga lexikografiska resurserna fanns tillgängliga som TBX (eller något annat vanligt förekommande format som används av översättningsprogram). Detta gäller särskilt översatta termer, eftersom översättningsprogrammen då inte bara kan hjälpa till vid själva översättningen, utan också göra en efterkontroll av att korrekta termer har använts.

I ett nordiskt perspektiv är översättning mellan typologiskt olika språk ingen nyhet, särskilt inte i Finland. Däremot innebär ett ökat fokus på minoritetsspråken att kunskap behöver byggas upp kring dessa. Därför bör den erfarenhet som redan finns, särskilt av översättning mellan germanska och uraliska språk, tas tillvara. Det kan också vara lämpligt att ta hjälp av typologer vid utformningen av lexikografiska resurser.

Det är fortfarande långt kvar till att de svenska nationella minoritetsspråken har god tillgång till språkteknologi och översättningsverktyg. Det som behövs är framförallt domänspecifika och stora lexikografiska resurser med en öppen licens samt stora textsamlingar.

Litteratur

Ordböcker och digitala resurser

Digital jiddisch–svensk–jiddisch ordbok. <sprak.isof.se/jiddisch/> (juni 2023).

Kerbel, Lennart, Jean Hessel & Peter David (2020): *Jiddisch–svensk–jiddisch ordbok*. Stockholm: Institutet för språk och folkminnen.

Lexin. <sprakresurser.isof.se/lexin/> (juni 2023).

Meänkieli–svensk–meänkieli ordbok. <sprak.isof.se/meankieli> (juni 2023), metadata: <snd.gu.se/sv/catalogue/study/2022-241> (juni 2023).

Myndighetstermlistan. <sprakresurser.isof.se/myndighetstermlistan/> (juni 2023).

IDS = Isofs digitala språkresurser. <sprakresurser.isof.se> (juni 2023).

Sverigefinska ordlistor. <snd.gu.se/sv/catalogue/collection/swedish---finnish-glossaries> (juni 2023).

Svensk nationell datatjänst. <snd.gu.se> (juni 2023).

Svensk-romska ordlistor. <isof.se/stod-och-sprakrad/spraktjanster/svensk-romska-ordlistor> (juni 2023).

Annan litteratur

Ahltorp, Magnus, Jean Hessel, Gunnar Eriksson & Maria Skeppstedt (2022): A Digital Swedish–Yiddish/Yiddish–Swedish Dictionary: A Web-Based Dictionary that is also Available Offline. I: *Proceedings of the EURALI Workshop @LREC2020*, 86–87.

Borin, Lars, Martha D. Brandt, Jens Edlund, Jonas Lindh & Mikael Parkvall (2012): *The Swedish Language in the Digital Age/Svenska språket i den digitala tidsåldern*. Berlin: Springer.

Dahl, Östen (2015): How WEIRD are WALS languages? I: *Diversity Linguistics: Retrospect and Prospect* (konferens 1 maj 2015). <www.eva.mpg.de/fileadmin/content_files/linguistics/conferences/2015-diversity-linguistics/Dahl_slides.pdf> (juni 2023).

Dahl, Östen (2019): Språk i skymundan. I: *Minoritetsspråk i Sverige och världen*. UPPLADOC, Vetenskap på kvällen (konferens 9 maj 2019).

ELE Consortium (2022): *Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap*. <european-language-equality.eu/wp-content/uploads/2022/11/ELE___Deliverable_D3_4__SRIIA_and_Roadmap___final_version_-1.pdf> (juni 2023).

Flerspråkig basterminologi som grund för tolkning och översättning (2022). Stockholm: Institutet för språk och folkminnen.

Fowler, Andrew, Kurt Partridge, Ciprian Chelba, Xiaojun Bi, Tom Ouyang & Shumin Zhai (2015): Effects of Language Modeling and its Personalization on Touchscreen Typing Performance. I: *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 649–658. doi:10.1145/2702123.2702503.

Nørstebø Moshagen, Sjur, Rickard Domeij, Kristine Eide, Peter Juel Henrichsen & Per Langgard (2022): *Report on the Nordic Minority Languages*. doi:10.1163/9789004298507.

Språklagen = Språklag (SFS 2009:600). <riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/sprakslag-2009600_sfs-2009-600/> (juni 2023).

Språklagen i praktiken – riktlinjer för tillämpning av språklagen (2011): Rapporter från Språkrådet 4. Stockholm: Institutet för språk och folkminnen.

UNESCO (2021): *World Atlas of Languages*. <en.wal.unesco.org> (maj 2023).

Marie Mattson
Språkvetare
Språkrådet, Institutet för språk och folkminnen
Alsnögatan 7
116 41 Stockholm
marie.mattson@isof.se

Magnus Ahltorp
Språkteknolog
Språkrådet, Institutet för språk och folkminnen
Alsnögatan 7
116 41 Stockholm
magnus.ahltorp@isof.se