Peter Widell och Ulf Dalvad Berthelsen (udg.):
11. Møde om Udforskningen af Dansk sprog
Århus 2006

Konditionell entropi som mått på lingvistiskt avstånd mellan danska och svenska

Av Jens Moberg & Charlotte Gooskens (Rijksuniversiteit Groningen)

Denna artikel beskriver ett projekt som bedrivs vid universitetet i Groningen. Projektet syftar till att använda konditionell entropi för att beräkna det lingvistiska avståndet[1] mellan de tre fastlandsspråken i Skandinavien (danska, norska och svenska). I denna artikel kommer endast språkparet danska–svenska att diskuteras. Genom att använda konditionell entropi vill vi undersöka vilken roll det lingvistiska avståndet spelar för den interskandinaviska språkförståelsen. Konditionell entropi kan också användas för att mäta asymmetrin i språkförståelse. Vi hoppas därmed få en inblick i huruvida lingvistiska faktorer är en del av förklaringen till varför danskar enligt flera perceptionsexperiment visat sig förstå svenska bättre än svenskar förstår danska.

1. Språksituationen i Skandinavien

Danska, norska och svenska är tre nära besläktade språk. Traditionellt har kommunikation mellan talare av dessa tre språk skett på respektive talares eget modersmål. Denna typ av kommunikation mellan ömsesidigt förståeliga språk har kommit att kallas semikommunikation (Haugen 1966). Ett flertal undersökningar, varav de viktigaste är Maurud (1976), Bø (1978) och Delsing & Lundin-Åkesson (2005), har ägnats detta fenomen och resultaten visar att möjligheten till semikommunikation varierar mellan språkparen. Bland annat uppvisar norrmän den bästa förståelsen av sina grannspråk, medan svenskar tycks förstå talad danska dåligt. En asymmetri i förståelse påvisas mellan svenska och danska, där danska är det språk som förstås sämre. De ovan nämnda undersökningarna föreslår att förklaringen till asymmetrin i den svensk-danska språkförståelsen inte fullt ut kan förklaras av lingvistiska faktorer.

I diskussionsdelen av Mauruds forskningsrapport kommenterar han skillnaden i förståelse mellan danska och svenska:

[1] Termen ”avstånd” används här även då det inte är lika långt från a till b som från b till a.

> "The low understanding of spoken Danish in Sweden can […] be explained by the contrast in sound structure between the two languages, but the Danes should have the same problems from this point of view. Thus we cannot explain from linguistic arguments alone why the Danes got a higher score on the test for spoken Swedish than the Swedes got on the test for spoken Danish."
>
> (Maurud 1976:69)

Maurud menar alltså att lingvistiska faktorer inte kan förklara asymmetrin mellan danska och svenska. Icke-lingvistiska faktorer såsom erfarenhet av det andra språket (via TV, släktingar semesterresor e.d.), geografisk närhet samt attityd tros ha stort inflytande. Maurud visade exempelvis att Sverige var det land som besöktes mest frekvent av andra skandinaver och att det var betydligt vanligare att norrmän och danskar såg på svensk TV än att svenskar såg TV-program från övriga Skandinavien. Delsing & Lundin-Åkesson frågade bland annat sina testpersoner vilket skandinaviskt språk de tyckte lät finast. Här var de svenska testpersonerna betydligt mer negativa till danska än vad de danska testpersonerna var till svenska.

Projektet som beskrivs i denna artikel har dock som syfte att objektivt utröna hur stort det rent lingvistiska avståndet är. Genom att jämföra detta avstånd med resultaten från ovan nämnda perceptionsexperiment hoppas vi få en inblick i vad som orsakar problemen i språkförståelsen.

En metod för att mäta lingvistiskt avstånd har nyligen testats vid universitetet i Groningen (Gooskens 2006). Detta gjordes genom att länka[2] fonetiskt transkriberade ordpar från de skandinaviska språken med varandra. Graden av fonetisk likhet bedömdes därefter med hjälp av den så kallade Levenshteinalgoritmen. Algoritmen går ut på att hitta den kortaste vägen mellan två strängar (i detta fall ordpar). Levenshteinalgoritmen beskrivs mer ingående i Heeringa (2004). Resultaten visar en hög korrelation mellan fonetisk likhet och förståelse. Denna korrelation är signifikant, till skillnad från de uppmätta sambanden mellan förståelse och extralingvistiska faktorer. Slutsatsen i undersökningen blir att det fonetiska avståndet i hög grad kan förutsäga förståelse men att mer precisa studier är önskvärda för att utforska relationen mellan fonetiskt (lingvistiskt) avstånd och språkförståelse. En av bristerna med metoden är att den inte kan uttrycka asymmetriska förhållanden. Avståndet från A till B är alltså alltid detsamma som från B till A, något som inte nödvändigtvis behöver vara fallet (se sektion 2.). Vår förhoppning är att konditionell entropi även ska kunna påvisa asymmetri i avståndet mellan danska och svenska.

[2] Med länkning menas en matchning av motsvarande segment så att det segment i ett språk som motsvarar ett segment i det andra språket hamnar intill varandra.

2. Entropi och konditionell entropi

Begreppet entropi[3] definierades av Shannon (1948) och betecknar graden av osäkerhet för en given sannolikhetsfördelning. Exempelvis kan man utifrån sannolikhetsfördelningen för att en kastad slant ska komma upp som krona beräkna entropin för denna möjlighet. Entropin kommer alltid vara som störst när sannolikheterna är jämnt fördelade, som i fallet med slantsingling där ju sannolikheten för krona är lika stor som sannolikheten för klave. I en lingvistisk kontext betyder detta att ju större valfrihet i val av enhet (bokstav, fonem e.d.) desto högre entropi. En låg entropi är ett tecken på liten osäkerhet, det vill säga liten valfrihet.

I konditionell entropi är variabel X en betingad sannolikhet av variabel Y. Den konditionella entropin för Y givet X uttrycker:

> ”how much extra information you still need to supply on average to communicate Y given that the other party knows X”
>
> (Manning & Schütze 1999:64).

Konditionell entropi är alltså ett mått på mängden information som fortfarande behöver tillföras Y givet vetskap om X. I detta projekt beräknas den konditionella entropin för en svensk talare som försöker hitta det svenska ljudsegment som motsvarar ett givet danskt ljudsegment och vice versa. Variablerna X och Y står här för ett ljudsegment i ett källspråk respektive ett målspråk.

Den praktiska tillämpningen av konditionell entropi inom lingvistik kan tydliggöras med ett exempel. Dagens skrivna danska har endast en vokal i grammatiska ändelser, nämligen *e*. Svenska har förutom *e* också *a* och *o*. Detta innebär att en svensk som påträffar den danska bokstaven *e* i en sådan kontext står inför tre val när denne skall hitta den svenska motsvarigheten. Den danska talare som påträffar den svenska bokstaven *a* eller *e* i en grammatisk ändelse vet däremot att detta alltid motsvarar danska *e*. Därför är entropin högre för svenska givet danska än tvärtom, i detta fall. Relationen är alltså asymmetrisk.

Ett utökat exempel illustrerar mer detaljerat hur den konditionella entropin kan räknas ut mellan två språk. Tabell 2.1 består av fyra ordpar:

Danska	**Svenska**
j ɑ i̯ (jeg)	j ɑː g (jag)
h a n (han)	h a nː (han)
v a ð (hvad)	v ɑː d (vad)
l ɑ ŋ ʔ (lang)	l ɔ ŋ # (lång)

Tabell 2.1. Fyra fonetiskt transkriberade dansk-svenska ordpar.

[3] Entropi används också som en term inom fysiken men här åsyftas informationsteoretisk entropi.

Ordparen består av tretton ljudsegment som länkas med varandra: [j] med [j], [ɑ] med [ɑː], [i] med [g] och så vidare. I det sista ordparet länkas dansk stød med en utfyllnadssymbol. I det här exemplet blir avståndet asymmetriskt på grund av den femte och elfte länkningen. I den femte länkningen matchas [a] med [a]. Svenska [a] förekommer bara på detta ställe och matchas därför bara med [a]. Danska [a] förekommer däremot på två ställen. Det matchas med [a] i den femte länkningen men med [ɑː] i den åttonde länkningen. Därför är sannolikheten för danskt [a] givet svenskt [a] 1 medan sannolikheten för svenskt [a] givet danskt [a] är 0.5. Samma sak sker med det danska segmentet [ɑ], som motsvaras av [ɑː] i den andra länkningen men med [ɔ] i den elfte länkningen. Dessa typer av korrespondenser orsakar asymmetri i det lingvistiska avståndet: valfriheten/osäkerheten blir större för den svenska talaren eftersom denne har fler ljudsegment att välja på än den danska talaren.

För samtliga länkningar samlas statistik in för sannolikheterna att dessa länkningar förekommer tillsammans. Utifrån dessa sannolikheter kan den konditionella entropin för svenska givet danska och danska givet svenska beräknas med en formel som vi av utrymmesskäl inte kommer att gå in på här (se Manning & Schütze 1999). Resultatet för de fyra ordparen i exemplet blir:

- 0,15 för danska givet svenska
- 0,31 för svenska givet danska

Entropin blir alltså högre för en svensk som försöker hitta den svenska motsvarigheten till ett givet danskt ljud. Om X är det givna segmentet så måste mer information tillföras när X är ett danskt ljudsegment för att få vetskap om Y (det motsvarande svenska ljudsegmentet) än tvärtom.

3. Material

För att kunna genomföra de olika entropiberäkningarna konstruerades en databas innehållande ordlistor från de tre skandinaviska språken. Databasen är också tänkt att användas för andra experiment och innehåller därför även tyska, frisiska och holländska. Dess innehåll utgörs av de mest frekventa orden från Corpus Gesproken Nederlands, CGN, (http://lands.let.kun.nl/cgn/home.htm, tillgänglig 061204) samt Europarl-korpusen (http://people.csail.mit.edu/koehn/publications/europarl, tillgänglig 061204). CGN är en

holländsk talspråkskorpus, ur vilken vi tog de 1 500 mest frekventa orden från den kategori som innehöll informellt tal. Denna del av korpusen innehåller totalt 2 626 172 icke-unika ord. Europarl är en talspråkskorpus som består av utdrag från Europaparlamentets sammanträden. Även där valdes de 1 500 mest frekventa orden ut från ett totalt urval på 889 836 icke-unika ord, denna gång från holländska och svenska. Europarl är översatt till många språk, anledningen till att vi valde att extrahera den holländska och svenska versionen var att dessa språk innehöll ordklassinformation, till skillnad från exempelvis danska. Dessutom representerar dessa språk de två germanska språkområden, nordgermanska och västgermanska, som databasen var tänkt att reflektera.

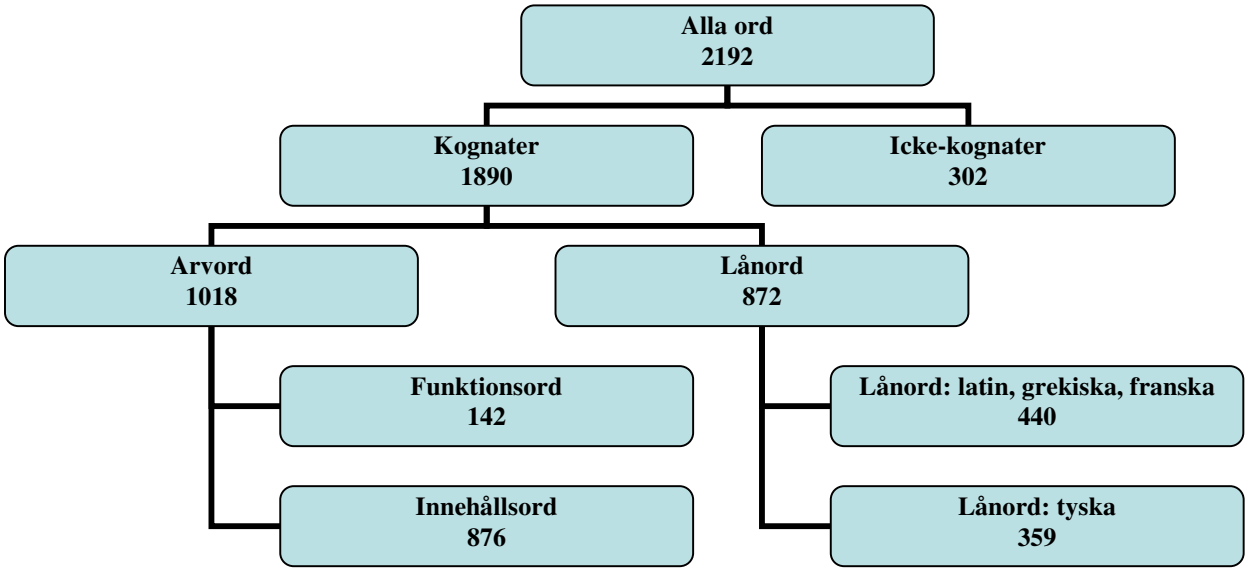
Databasen kompletterades med funktionsord hämtade från diverse grammatikor, detta i syfte att få en så heltäckande samling funktionsord som möjligt. Egennamn samt interjektioner avlägsnades. Vår motivering till att basera entropiberäkningarna på dels formella, dels informella ord är att vi förväntade oss en skillnad i antalet lånord i dessa två kategorier. Europarl innehåller många ord från de politiska och ekonomiska domänerna. Dessa ord är ofta lånade från latin eller grekiska. CGN:s informella tal är däremot hämtat från vardagliga talsituationer där vi förväntar oss fler arvord. Dessa har haft längre tid att utvecklas och modifieras enligt språkets ljudutveckling, vilket kan förväntas ge större avvikelse från det relaterade språket och därmed uppvisa en skillnad i entropi jämfört med Europarls lånord.

Materialet översattes samt transkriberades efter standardiserat uttal som det angetts i uttalslexikon. Den slutliga versionen av databasen innehåller följande information per ord och språk:

- Vilken korpus ordet hämtats från
- Ordklass
- Funktionsord/innehållsord
- Ordets ursprung (lånord eller arvord)
- Om ordet är ett lånord anges vilket språk ordet lånats in från
- Ordets lexikala representation
- Ordets fonetiska representation
- Kognat/icke-kognat (om ett ord från ett språk är kognat till ett ord i ett annat språk har orden samma siffra i kolumnen Kognat)

Termen *kognat* (eng. *cognate*) brukar beskrivas som två ord från olika språk som har ett gemensamt ursprung och där orden är arvord i sina respektive språk. I vår undersökning används dock termen i en utvidgad betydelse: för ordpar där orden är genetiskt relaterade, d.v.s. delar ett gemensamt ursprung, antingen genom att de båda är arvord eller för att de härstammar från samma språk. I två så nära besläktade språk som danska och svenska är den stora majoriteten ordpar kognater, exempelvis *gammel–gammal* och *direktør–direktör*. Exempel på icke-kognater är *kun–endast*.

Vi valde att dela upp databasen i ett antal kategorier för att se hur stor skillnaden i entropi blev när olika typer av ord testades (se sektion 4). Dessa kategorier samt deras storlek framgår av trädet i figur 3.1. Den grammatiska indelningen i funktionsord/innehållsord gjordes endast för arvord eftersom i stort sett samtliga funktionsord återfinns i den kategorin.



Figur 3.1. Databasens innehåll och storlek.

4. Hypotes

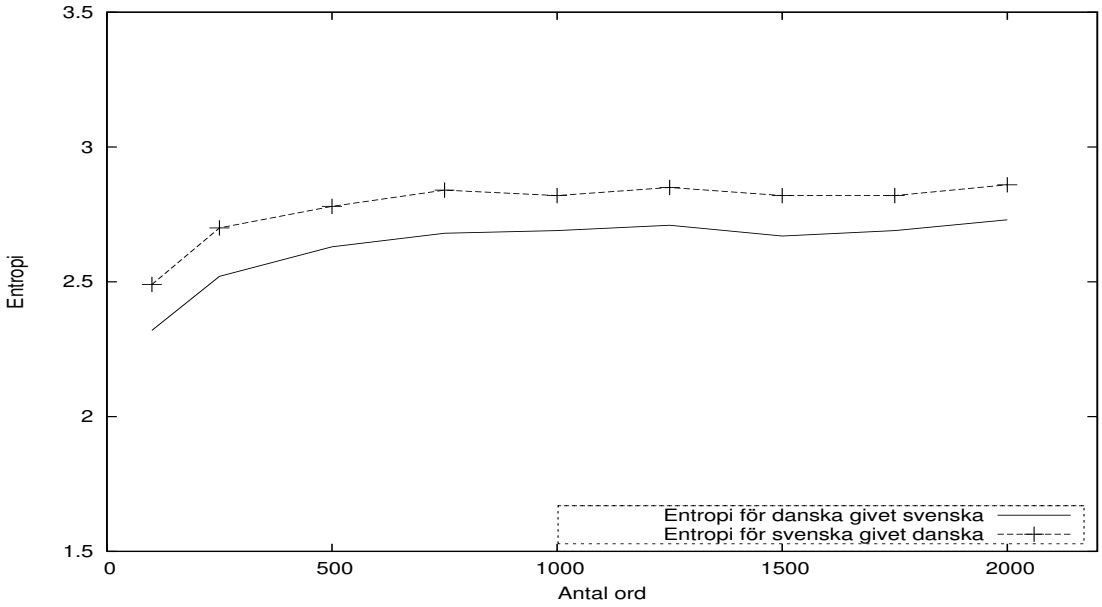
Det vi hoppades få bekräftat var framförallt att entropin för svenska givet danska skulle vara högre än den för danska givet svenska, alltså att det lingvistiska avståndet skulle vara asymmetriskt.

Vi förväntade oss högst entropi för den översta noden i figur 3.1, bestående av samtliga ord. Då även icke-kognater ingår i denna grupp lär andelen regelbundna korrespondenser nämligen vara lägre än för bara kognater. Vi hade även en hypotes om att arvorden skulle uppvisa större

skillnader mellan språken än lånord, eftersom de har utvecklats i respektive språk under lång tid. Lånord som kommit in i språket förväntas vara mer lika då de sannolikt inlånats i liknande form och sedan inte hunnit divergera lika mycket. Uppdelningen i dels romanska och grekiska lånord, dels tyska lånord gjordes eftersom vi misstänkte en skillnad i entropi då bland annat tiden för inlåning för dessa båda kategorier skiljer sig åt. Den stora inströmningen av lånord från tyska till svenska och danska skedde framförallt på 1300- och 1400-talet, under Hansans storhetstid. På 1700-talet blev det däremot populärt att importera franska ord (Edlund & Hene 1992). De ord som importerats till svenska respektive danska har möjligen anpassats på ett regelbundet sätt till respektive språks ljudsystem. Ett inlånat ord som innehåller ljudet X har blivit det svenska ljudet Y och det danska ljudet Z så att relationen mellan Y och Z är regelbunden. Då de ursprungligen tyska orden funnits i svenska och danska längre än de franska, har de under längre tid utvecklats enligt det lokala ljudsystemet vilket skulle kunna medföra att regelbundenheterna uppluckras. Tiden för inlåning skulle därmed kunna innebära att kategorin romanska samt grekiska lånord uppvisar lägre entropi.

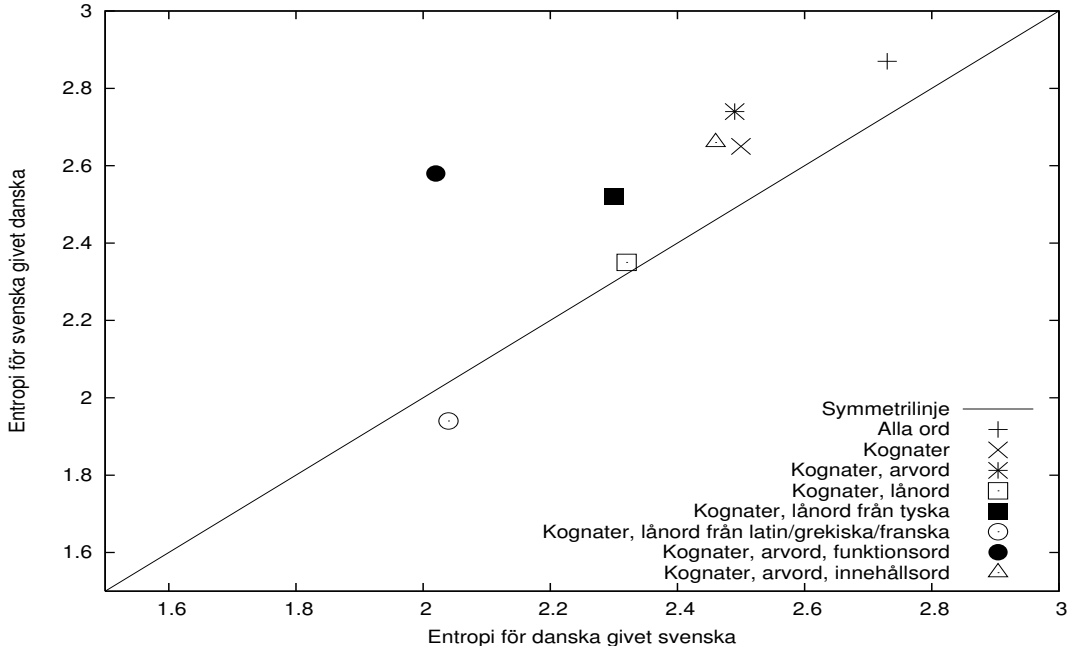
5. Resultat

Det första vi ville ta reda på var hur mycket antalet ord påverkar entropiresultaten. Eftersom våra kategorier är ojämnt fördelade med avseende på detta (se figur 3.1) var ett sådant test nödvändigt. Figur 5.1 visar hur den konditionella entropin påverkas av antalet ord.



Figur 5.1. Förhållandet mellan konditionell entropi och antalet ordpar.

Diagrammet baseras på 2000 slumpmässigt utvalda[4] dansk-svenska ordpar. Dessa ordpar delades upp i nio grupper, innehållande från 100 till 2000 ordpar. Genom att räkna ut den konditionella entropin för varje grupp kan man studera hur kurvorna ändras allteftersom ordantalet stiger. De två kurvorna har en tämligen likartad utveckling, vilket tyder på att asymmetrin i entropi inte påverkas så mycket av antalet ord. Det största hacket i kurvan dyker upp efter cirka 250 ord men därefter sker en utplaning och entropin håller sig på en relativt konstant nivå.



Figur 5.2. Konditionell entropi mellan danska och svenska för hela databasen.

Figur 5.2 visar hur asymmetriska de testade kategorierna är samt hur hög entropin är för varje enskild kategori. X-axeln visar entropin för danska givet svenska, alltså hur svårt det är för en dansk att förutsäga den danska motsvarigheten till ett givet svenskt ljudsegment. Y-axeln visar entropin för svenska givet danska. Den heldragna linjen symboliserar en helt symmetrisk situation, där entropin är lika hög i båda riktningarna. De punkter som är ovanför linjen uppvisar asymmetri i form av en högre entropi för svenska givet danska (större valfrihet/osäkerhet för en svensk) medan punkter under linjen har högre entropi för danska givet svenska. Samtliga testade kategorier rör sig i intervallet mellan 2 och 3 bits[5]. Kategorin innehållande samtliga ord uppvisar den högsta entropin. Förklaringen är sannolikt en kombination av att det är den största kategorin

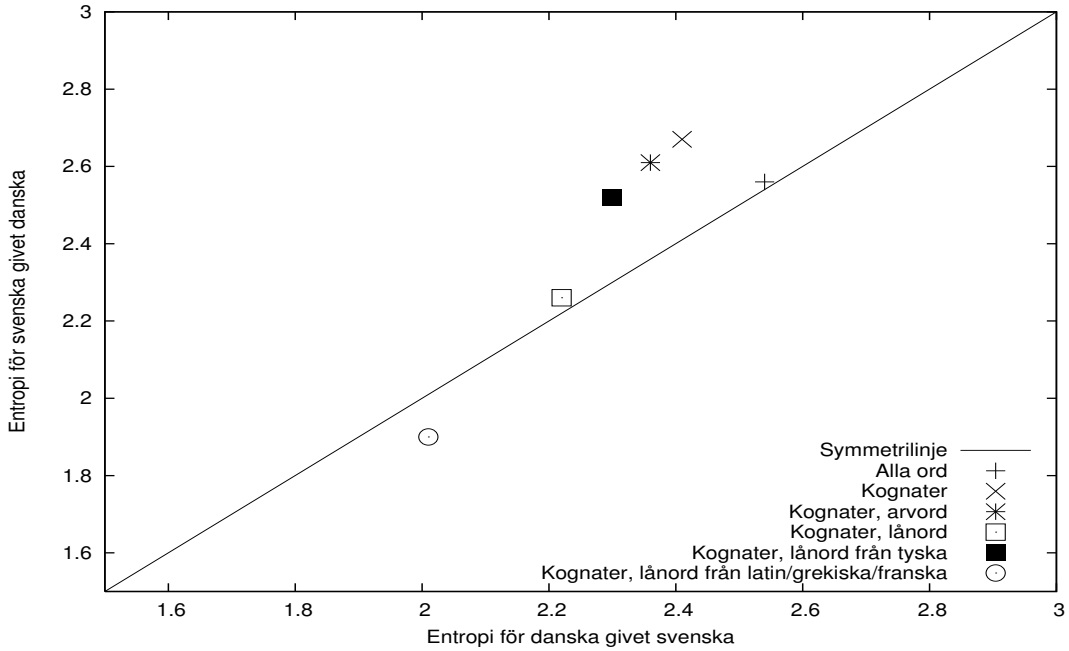
[4] Urvalet gjordes med en slumpgenerator som finns tillgänglig online: http://www.randomizer.org, tillgänglig 22/11-06.

[5] Bits är den enhet som entropi mäts i. Bits står för binary digits, alltså binära enheter. Dessa uttrycker ett val mellan två alternativ.

(se figur 5.1) samt att den innehåller ett antal icke-kognater. Det visar sig också att kategorin arvord hamnar högre än kategorin lånord, något som stämmer väl överens med vår hypotes (se sektion 4).

Kategorin kognater bestående av romanska och grekiska lånord är den enda kategori som avviker från trenden och istället uppvisar högre entropi för danska givet svenska. Kategorin ger också den lägsta entropin i båda riktningarna, vilket överensstämmer med hypotesen. Kategorin funktionsord avviker relativt kraftigt från symmetrilinjen men denna kategori består endast av ca 150 ord och några säkra slutsatser kan därför inte dras.

För att kunna slå fast hur stor betydelse antalet ord har gjordes nya beräkningar, där samtliga kategorier innehåller lika många ord. Med utgångspunkt från den minsta kategorin, tyska lånord (kategorin funktionsord valdes bort eftersom den ansågs för liten) valdes 344 ordpar slumpvis ut från de övriga kategorierna, samma antal ordpar som i kategorin tyska lånord. Därefter sattes resultaten in i ett diagram på samma format som ovanstående, men i detta fall är kategorierna alltså lika stora.



Figur 5.3. Konditionell entropi mellan danska och svenska baserat på 344 ordpar per kategori.

De inbördes relationerna mellan kategorierna är ungefär desamma. Lånord från latin, grekiska och franska ger fortfarande en högre entropi för danska givet svenska medan de övriga kategorierna uppvisar högre entropi för svenska givet danska. En jämförelse mellan figur 5.2 och

5.3 samt tendensen som uppvisas i figur 5.1 ger en sammantagen bild av att antalet ordpar endast har en marginell inverkan på resultatet så länge antalet överstiger 250.

6. Sammanfattning

På basis av en databas bestående av formellt och informellt tal har vi beräknat den fonetiska delen av det lingvistiska avståndet mellan danska och svenska. Detta har gjorts med en formel för konditionell entropi. Entropin mellan danska och svenska har räknats ut på fonetisk nivå för ett antal kategorier innehållande ord från olika språk. Resultaten stämmer väl överens med den på förhand givna hypotesen, liksom det faktum att entropin överlag är högre för svenskar som försöker förstå danska än för danskar som försöker förstå svenska. Det ser alltså ut som om lingvistiska faktorer kan vara en del av förklaringen till den asymmetriska dansk-svenska språkförståelsen.

Vi har även sett att ord som lånats in från latin, franska och grekiska uppvisar en låg entropi i båda riktningarna: svenska givet danska och danska givet svenska. Detta är speciellt intressant att kontrastera med kategorin lånord från tyska. Där är entropin betydligt högre i båda riktningarna.

Litteratur

Bø, I. (1978). *Ungdom og naboland*. Stavanger: Rogalandsforskning (rapport 4).

Delsing, L-O och Lundin-Åkesson, K. (2005). *Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska.* Köpenhamn: Nordiska ministerrådet.

Edlund, L-E och Hene, B. (1992). *Lånord i svenskan – om språkförändringar i tid och rum.* Umeå och Stockholm: Förlags AB Wiken.

Gooskens, C. (2006). Linguistic and extra-linguistic predictors of Inter-Scandinavian intelligibility. I: Jeroen van de Weijer & and Bettelou Los (eds.). *Linguistics in the Netherlands* 23, Amsterdam: John Benjamins:101-113.

Haugen, E. (1966). Semicommunication: The language gap in Scandinavia. *Sociological inquiry* 36:280-297.

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* Groningen: Groningen dissertations in linguistics (Grodil).

Manning, C. och Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press.

Maurud, Ø. (1976). Reciprocal Comprehension of Neighbour Languages in Scandinavia. An investigation of how well people in Denmark, Norway and Sweden understand each other's written and spoken languages. In: *Scandinavian journal of educational research* 20:49-72.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal* 30:50-64.