EN STUDIE ÖVER PÅVERKANDE FAKTORER I ORDKLASS-TAGGNING. BASERAD PÅ TAGGNING AV SVENSK TEXT MED EPOS.

Sofie Johansson Kokkinakis



EN STUDIE ÖVER PÅVERKANDE FAKTORER I ORDKLASS-TAGGNING. BASERAD PÅ TAGGNING AV SVENSK TEXT MED EPOS.

(English Summary)

Sofie Johansson Kokkinakis



Akademisk avhandling för filosofie doktorsexamen i språkvetenskaplig databehandling, som enligt beslut av humanistiska fakultetsnämnden vid Göteborgs universitet kommer att försvaras offentligt lördagen den 18 januari 2002 kl. 10 i sal T 302, Arkeologen.

Abstract

Johansson Kokkinakis, Sofie, *En studie över påverkande faktorer i ordklasstaggning. Baserad på taggning av svensk text med EPOS.* Göteborg: 2002, 221 s. Språk: svenska. Doktorsavhandling i språkvetenskaplig databehandling. Institutionen för svenska språket, Göteborgs universitet. *A study of factors influencing performance in part of speech tagging. Based on tagging of Swedish text with EPOS.* Language: Swedish. Doctoral dissertation in Natural Language Processing. Department of Swedish Language, Göteborg University. 2002. 221 pp.

Part of speech (POS) tagging is known to be one of the first steps in natural language processing. The aims of the study is to determine to what degree it is possible to enhance performance in POS tagging and at what cost. These issues are addressed by examining five different factors bearing on the performance of part of speech tagging in Swedish. These factors are divided into two subgroups depending on how directly or indirectly they seem to affect performance. The *direct* factors are lexical as well as textual initial text material and computational method, whereas the *indirect* ones are tagset and evalutation method.

In order to determine the most adequate approach to improving performance in POS tagging, results from the statistical method were compared with approaches focusing on well-established problem areas. It is concluded that statistical methods are certainly very important, up to a point, when it comes to correctly tagged text. In the end, however the approaches which focused on specific problem areas seemed to have greater effect.

Special attention was paid to evaluation. The performance of the tagger used was tested on a text of 100,000 words, and the errors discovered (5,000) were examined in detail. Since ordinary evaluation measures seem far too superficial to account for the relevant factors underlying the errors, a new evaluation method is proposed, $H_{measure}$, in order to handle homography in the materials more adequately.

Key words: part of speech tagging, lexical databases, ambiguity, homography, disambiguation, tagsets, training and test corpora, machine learning, evaluation.

© Sofie Johansson Kokkinakis 2002

Sofie Johansson Kokkinakis

EN STUDIE ÖVER PÅVERKANDE FAKTORER I ORDKLASSTAGGNING. BASERAD PÅ TAGGNING AV SVENSK TEXT MED EPOS.

Abstract

Johansson Kokkinakis, Sofie, *En studie över påverkande faktorer i ordklasstaggning. Baserad på taggning av svensk text med EPOS.* Göteborg: 2002, 221 s. Språk: svenska. Doktorsavhandling i språkvetenskaplig databehandling. Institutionen för svenska språket, Göteborgs universitet. *A study of factors influencing performance in part of speech tagging. Based on tagging of Swedish text with EPOS.* Language: Swedish. Doctoral dissertation in Natural Language Processing. Department of Swedish Language, Göteborg University. 2002. 221 pp.

Part of speech (POS) tagging is known to be one of the first steps in natural language processing. The aims of the study is to determine to what degree it is possible to enhance performance in POS tagging and at what cost. These issues are addressed by examining five different factors bearing on the performance of part of speech tagging in Swedish. These factors are divided into two subgroups depending on how directly or indirectly they seem to affect performance. The *direct* factors are lexical as well as textual initial text material and computational method, whereas the *indirect* ones are tagset and evalutation method.

In order to determine the most adequate approach to improving performance in POS tagging, results from the statistical method were compared with approaches focusing on well-established problem areas. It is concluded that statistical methods are certainly very important, up to a point, when it comes to correctly tagged text. In the end, however the approaches which focused on specific problem areas seemed to have greater effect.

Special attention was paid to evaluation. The performance of the tagger used was tested on a text of 100,000 words, and the errors discovered (5,000) were examined in detail. Since ordinary evaluation measures seem far too superficial to account for the relevant factors underlying the errors, a new evaluation method is proposed, $H_{measure}$, in order to handle homography in the materials more adequately.

Key words: part of speech tagging, lexical databases, ambiguity, homography, disambiguation, tagsets, training and test corpora, machine learning, evaluation.

© Sofie Johansson Kokkinakis 2002

Distribution: Inst. för svenska språket (Dept. of Swedish Language)
Box 200, SE-405 30 Göteborg, Sweden

Förord

Anledningen till att jag när jag var färdig med min magisterexamen ansökte till forskarutbildningen i språkvetenskaplig databehandling var att dåvarande linjeledaren vid datalingvistprogrammet Gudrún Magnúsdóttir övertygade mig om att jag måste besöka hennes avdelning Språkdata vid institutionen för svenska språket. Det gjorde jag och erbjöds så småningom arbete som systemman där. Det förväntades också att man skulle doktorera. Eftersom jag inte hade några andra konkreta planer påbörjade jag forskarutbildningen som fritidsdoktorand 1992. Därefter läste jag kurser samtidigt som jag arbetade heltid. Heltidsdoktorand blev jag först 1998, fast egentligen inte på heltid heller. Under de drygt fyra åren som jag officiellt varit doktorand har jag haft försörjning i form av doktorandtjänst från humanistiska fakulteten under 2 år och 8 månader. Övrig tid har jag inte kunnat arbeta på avhandlingen utan varit mammaledig eller arbetat i projekt. Avhandlingsarbetet kan därefter fördelas på två tidsperioder där den första delen (ca 80%) av tiden har ägnats åt datortekniskt utvecklingsarbete och utvärdering, och den andra delen (ca 20%) har gått till teoretiska studier och dokumentation av det arbete som manifesteras i den tryckta avhandlingen.

Det finns olika inspiratörer till min avhandling. Mitt intresse för språk får jag tillskriva bl.a. Astrid Lindgren för hennes figur Pippi som ägnar mycket tid åt att ta reda på vad ordet "spunk" betyder, vilket enligt Pippi är "ett ord som profissorerna inte har uppfunnit ännu". Det visar sig till slut vara en skalbagge. Även Sven Nordqvist har genom sin Pettson och Findus bidragit till en del finurligheter såsom en stor ökning av ord som börjar på "o-", "min köttbulle o-finns", "det o-ligger potatis i vartenda hål", "o-hönsspår överallt", "o-sov så gott". Ett annat mer generellt språkvetenskapligt ämne är den särskrivning som blir allt vanligare i svenskan. Det har dock tillfört en del livliga diskussioner och många skratt vid kaffebordet då det är ett ämne som engagerar de flesta kollegor vid vår avdelning. Ta t.ex. Toys 'r us reklam "sökes kassa medarbetare..." eller från en kontaktannons "kort hårig sjuk sköterska", "rök fritt" eller frasen "är du djur och naturintresserad?". Ämnet homografi stämmer in på denna beskrivning. Ingen slår givetvis Sture Berg då han slänger ur sig tvetydigheter vid kaffebordet. Homografi och ordförvrängande är något som märks alltmer i samhället, t.ex. i reklam. "En svensk tiger" är nog det mest klassiska exemplet på homografa fraser. Sedan finns det andra exempel på ordförvrängande t.ex. i Sibas reklam "nu är det jul-i igen". Som språkvetare är det en fröjd att se mina egna barns uppfinningsrikedom när det gäller ord. "Nummerdator" (miniräknare) och "min böck" (bok) är några exempel på detta. Som 4-åring uppfann jag ordet "stabant" (stabil+

bastant).

Följande personer har på olika sätt har gett mig stöd i mitt avhandlingsarbete.

Först ett stort tack till min huvudhandledare Jerker Järborg som haft tålamod med mig då jag t.ex. har ändrat inriktning på avhandlingen minst två gånger, knackat på hans dörr två gånger i kvarten, men som ändå alltid trott på mina idéer och varit ett mycket positivt stöd i alla lägen. Tack också Lars Borin för ett utmärkt opponentskap vid mitt slutseminarium och för litteraturtips och initierande synpunkter i slutskedet. Lars Ahrenberg för hans bihandledarskap fram till 2001 och hans konstruktiva idéer.

Yvonne Cederholm för tillgång till SMDB och hjälp med framtagning av underlag för beskrivning av databasen. Sture Berg och Martin Gellerstam för deras synpunkter om homografi och korrekturläsning av vissa utvalda delar. Sofia Gustafson-Capkova för tillgång till senaste versionen av SUC och beskrivning därom. Sölve Ohlander för hans hjälp med språkgranskning av den engelska delen av avhandlingen. Urban Hjort för granskning av statistiska formler och idéer om nya undersökningar. Torbjörn Lager för konsultation i ett tidigt skede om datatekniska metoder. Beáta Megyesi för givande diskussioner om faktorer som påverkar prestation i POS-taggning. Dimitris Kokkinakis för litteraturtips och korrekturläsning. Elisabet Engdahl för värdefulla råd och tips om avhandlingsskrivande i allmänhet.

Bland utländsk expertis har jag konsulterat följande personer: Steven Abney om grammatisk parsning och SCOL-systemet vid utveckling av grammatisk parsning som följer arbetet med ordklasstaggning. David Elworthy om utvärdering av ordklasstaggning av svenska och studier om tagguppsättningar och korpusars betydelse för kvalitet i ordklasstaggning. Fred Karlsson för en givande diskussion om homografins betydelse och orsaker till att den mäts på olika sätt. Dan Tufis för hans insikter om tagguppsättningars betydelse för kvalitet i ordklasstaggning.

Jag vill också tacka alla arbetskamrater, släkt och vänner som på något sätt har varit inblandade i mitt avhandlingsarbete, alla som varit stöttande i svåra stunder på arbetet och privat. Speciellt tack till min kompis Susanne Mankner som alltid kan se något positivt även då det ser som mörkast ut, och Gunnar Dahlström utan vars hjälp krånglande datorer och datorprogram hade varit oöverstigliga hinder.

Mina föräldrar, Airi Rovio-Johansson och Roger Johansson, vill jag tacka för deras hjälp med konstruktiva synpunkter på olika textavsnitt och

allmänt stöd hemma.

Mina döttrar Emelie och Josefin förtjänar ett speciellt tack för att de har varit duktiga töser och länge accepterat att jag har arbetat på kvällar och helger. Sist men inte minst Dimitris för att han har ställt upp på hemmafronten när jag arbetat eller varit trött.

Göteborg, december 2002

Sofie Johansson Kokkinakis

Innehållsförteckning

1 **Inledning** 1
1.1 Bakgrund 1
1.2 Syfte 2
1.3 Tillvägagångssätt i avhandlingsarbetet 3
1.3.1 Faktorer som påverkar prestationen i taggning 3
1.3.2 Förbättring av prestation 5
1.3.3 Bedömning av prestation 6
1.3.4 Taggningsverktyget EPOS 7
1.4 Avhandlingens uppläggning 8

2 **Forskningsöversikt** 11
2.1 Direkta faktorer 11
2.1.1 Lexikaliskt material 12
2.1.1.1 Enhetsförråd 12
2.1.1.2 Informationsstruktur 15
2.1.2 Textuellt material 19
2.1.2.1 Kvalitet och innehåll 19
2.1.2.2 Storlek 22
2.1.3 Disambigueringsmetod 22
2.1.3.1 Inledning om ordklasstaggning 23
2.1.3.2 Metoder och tekniker vid disambiguering 28
2.2 Indirekta faktorer 41
2.2.1 Tagguppsättning - standarder och rekommendationer 41
2.2.2 Utvärderingsmetod 45
2.2.2.1 Vad är korrekt taggning? 45
2.2.2.2 Bakgrund om termer och begrepp inom utvärdering 48
2.2.2.3 Definition av utvärderingsmått 49

3 **Lexikaliska och textuella material** 55
3.1 Lexikon 55
3.1.1 SAOL-SMDB 55
3.1.2 NFO 57
3.2 Korpusar 58
3.2.1 Press 65 58
3.2.2 SUC 60

4 **EPOS: utvecklings- och metodavsnitt** 63
4.1 Mappningsmetod 63
4.1.1 Jämförelser av olika tagguppsättningar 64
4.1.2 Modulen "Suc2saol" 70
4.1.3 Skillnader och problem 81
4.1.3.1 Skillnader i tagguppsättningar 81
4.1.3.2 Problem i mappningen 89
4.2 Disambigueringsmetod - EPOS-En Probabilistisk Ordklasstaggare för Svenska 96
4.2.1 Bakgrund 96
4.2.2 EPOS-modellen 97
4.2.2.1 Design och innehåll 97
4.2.2.2 Machine Learning vid konstruktion av EPOS 102
4.2.2.3 Tokenisering 105
4.2.2.4 Taggning med hjälp av lookup-metod och fullformsdatabas 106
4.2.2.5 Morfologisk analys av okända ord 107
4.2.2.6 Disambiguering med hjälp av trigram, token och sannolikhet 114
4.2.2.7 Gränssnitt och kommunikation 123
4.2.3 Tillvägagångssätt och problem i disambigueringen 128

5 **Utvärdering** 131

5.1 Två utvärderingsmått som tar hänsyn till vissa faktorer 131

5.1.1 Hmeasure 132

5.2 Diskussion om vidare undersökning av Hmeasure 136

5.3 Jämförelse med andra utvärderingsmått 139

5.4 Utvärderingsverktyg 139

5.5 Resultat av olika utvärderingar 142

5.5.1 Prestation indelad i ordklasser 143

5.5.2 Analys av fel 147

5.6 Slutsatser 157

6 **Diskussion om faktorer som påverkar prestation** 159

6.1 Direkta faktorer 159

6.1.1 Träningsmaterial 159

6.1.2 Lexikaliskt material 162

6.1.3 Disambigueringsmetod 163

6.2 Indirekta faktorer 167

6.2.1 Taggguppsättning 167

6.2.2 Utvärderingsmetod 171

6.3 Slutsatser om påverkande faktorer 171

7 **Diskussion om undersökning av förbättringsmöjligheter** 173

7.1 Förbättring av mappat träningsmaterial 173

7.2 Förbättring av disambigueringen 181

8 **Slutsatser** 189

8.1 Slutsatser om direkta faktorer 189

8.1.1 Textuellt utgångsmaterial 189

8.1.2 Lexikaliskt utgångsmaterial 191

8.1.3 Metod 191

8.2 Slutsatser om indirekta faktorer 192

8.2.1 Taggguppsättning 192

8.2.2 Utvärderingsmetod 193

8.3 Vidare forskning 194

8.4 Sammanfattning 195

9 **Summary** 197

9.1 Purpose 198

9.2 Conclusions 199

9.2.1 Direct factors 199

9.2.1.1 Initial textual material 199

9.2.1.2 Initial lexical material 201

9.2.1.3 Method 201

9.2.2 Indirect factors 202

9.2.2.1 Tagset composition 202

9.2.2.2 Evaluation method 203

9.3 Final conclusions 204

Litteraturreferenser 205

APPENDIX A. Taggguppsättning för EPOS 1.0 213

APPENDIX B. Taggguppsättning för SMDB/EPOS 2.0 217

Figurförteckning

Figur 1. Beskrivningsnivåer från NFO, Allén (1970). 16
Figur 2. En homografkomponents uppbyggnad i NFO, Allén (1970). 18
Figur 3. Generell beskrivning av taggningssystem, McEnery och Wilson (1997:121). 26
Figur 4. Generell modell av ett inducerat inlärningssystem, Daelemans (1999:288). 31
Figur 5. Systemet runt SMDB, från Berg, Cederholm & Gellerstam (2003). 57
Figur 6. Modul för mappning av SUC:s taggуppsättning till SMDB:s taggуppsättning. 70
Figur 7. Gränssnitt till konkordansverktyg utvecklat inom avhandlingen för EPOS 1.0. 90
Figur 8. Gränssnitt till konkordansverktyg utvecklat inom avhandlingen för EPOS 1.0. 91
Figur 9. EPOS-modellen. 97
Figur 10. Beskrivning av ett inducerat inlärningssystem i EPOS, Daelemans (1999). 104
Figur 11. Uppdelning av en taggad mening i trigram i träningen av EPOS. 104
Figur 12. Metod för analys av okända ord i den morfologiska modulen i EPOS. 111
Figur 13. Antal möjliga kombinationer av taggar, med sannolikhetsvärden från EPOS. 112
Figur 14. Exempel på analys med trigram enligt mönstret: n-1, n, n+1 i EPOS. 118
Figur 15. Exempel på analys med trigram enligt mönstret: n-2, n-1, n. 119
Figur 16. Beskrivning av en del av gränssnittet i EPOS v.1.0. 124
Figur 17. Beskrivning av gränssnittet efter en körning i EPOS. 125
Figur 18. Gränssnitt till sökning på ord och taggar i EPOS 1.0. 125
Figur 19. Gränssnitt (utvecklat inom avhandlingen) till resultat från sökning på taggar i EPOS 1.0. 126
Figur 20. Sökgränssnitt till Språkbankens korpusbearbetningsverktyg. 126
Figur 21. Resultat i sökgränssnitt till Språkbankens korpusbearbetningsverktyg. 127
Figur 22. Förhållandet mellan medelvärdet av $H_{measure}$ i lexikon (H_{lex}) och text (H_{text}), med värden från EPOS 2.0. 135
Figur 23. Hmeasure och recall med procent korrekta taggar och antal taggar per ord i EPOS. 136
Figur 24. Uppskattning av prestation i förhållande till en viss svårighet. Homografi i lexikonet. Från EPOS. 137
Figur 25. Uppskattning av prestation i förhållande till en viss svårighet. Homografi i texten. Från EPOS. 138
Figur 26. EUFOROS och modulerna i utvärderingsverktyget utvecklat inom avhandlingen. 141

Figur 27. Spridning av prestation vid disambiguering i form av punktdiagram från utvärdering av EPOS 2.0 och 1.0. 146
Figur 28. Spridning av prestation vid analys av okända ord i form av punktdiagram från utvärdering av EPOS 2.0 och 1.0 146
Figur 29. Cirkeldiagram över fördelning av disambigueringsfel i EPOS 2.0.... 150
Figur 30. Diagram över förhållandet mellan prestation och storlek på träningskorpusar, Megyesi (2002). 161
Figur 31. Inlärningskurva vid träning av TnT-tagger på NEGRA-korpusen, Brants (1998). 162
Figur 32. Diagram vilket illustrerar relation mellan prestation och taggуppsättningsstorlek, Megyesi (2001). 168

Tabellförteckning

Tabell 1. Exempel på homografseparering och lemmatisering, Allén S. NFO (1970). 17

Tabell 2. Fördelning av graford, homografkomponenter och lemman från tabell 1. 17

Tabell 3. Exempel på representation av indata i Machine Learning, Daelemans (1999:288). 31

Tabell 4. Exempel på representation av utdata i Machine Learning, Daelemans (1999:289). 32

Tabell 5. Alternativ vid intern representation, i Machine Learning, Daelemans (1999:289). 32

Tabell 6. Förenkling av alternativ vid ökad inlärning/kunskapstillväxt, i Machine Learning, Daelemans (1999:289). 32

Tabell 7. Exempel på användningssätt vid "Table look-up" metoden, i Machine Learning, Daelemans (1999:290). 33

Tabell 8. Exempel på användningssätt vid "Case-Based Learning" metoden, i Machine Learning, Daelemans (1999:290). 33

Tabell 9. Exempel på användningssätt vid "Rule and Decision Tree Induction", i Machine Learning, Daelemans (1999:291). 34

Tabell 10. Obligatoriska attribut och värden, ur EAGLES rekommendationer (1999). 43

Tabell 11. Rekommenderade attribut och värden, ur EAGLES rekommendationer (1999). 43

Tabell 12. Speciella extensioner. Valfria språkspecifika attribut och värden, ur EAGLES rekommendationer (1999). 43

Tabell 13. Exempel "gamla testamentet" från SUC 1.0, 2.0 och SMDB med olika taggning. 47

Tabell 14. Exempel på skillnad i taggning mellan SUC och SMDB. 65

Tabell 15. Kategorisering av proprier i SUC, Källgren (1998:31). 71

Tabell 16. Exemplet "på andra håll" från taggning med SMDB och EPOS. 73

Tabell 17. Exemplet " ... att Landsbergis kompromisslösa, stundom utmanande, ...", från taggning med SMDB och EPOS. 74

Tabell 18. Exempel "Don Kerr, en av de politiska tänkarna ... ". 76

Tabell 19. Fördelning av taggar i SUC:s och SMDB:s tagguppsättningar. 81

Tabell 20. Skillnader vid artikel, exempel från mappningsarbetet mellan SUC och SMDB. 82

Tabell 21. Skillnader vid räkneord, exempel från mappningsarbetet mellan SUC och SMDB. 82

Tabell 22. Skillnader vid nomen, exempel från mappningsarbetet mellan SUC och SMDB. 83

Tabell 23. Skillnader vid egennamn, exempel från mappningsarbetet mellan SUC och SMDB. 84

Tabell 24. Skillnader vid pronomen, exempel från mappningsarbetet mellan SUC och SMDB. 85

Tabell 25. Skillnader vid adverb, exempel från mappningsarbetet mellan SUC och SMDB. 85
Tabell 26. Skillnader vid verb, exempel från mappningsarbetet mellan SUC och SMDB. 86
Tabell 27. Skillnader vid förkortningar, exempel från mappningsarbetet mellan SUC och SMDB. 87
Tabell 28. Exempel på ord som skiljer sig i taggning mellan SUC och SMDB. .. 92
Tabell 29. Exempeldata i EPOS-modellen. 98
Tabell 30. Ett lemma och dess böjningsformer, från SMDB. 106
Tabell 31. Beskrivning av två modeller för trigram vid morfologisk analys. 112
Tabell 32. Exempel på tre regler med trigram i EPOS, plus två andra regler vilka används vid disambiguering. 121
Tabell 33. Kategorifel och morfologiska fel i EPOS 2.0 139
Tabell 34. Prestationsinformation från EPOS 2.0 och EPOS 1.0. 142
Tabell 35. Information om disambiguering i EPOS 2.0 och 1.0. 142
Tabell 36. Information om analys av okända ord i EPOS 2.0 och 1.0. 143
Tabell 37. Prestation med avseende på felindelning i EPOS 2.0 och 1.0. 143
Tabell 38. Prestation med avseende på ordklassindelning vid disambiguering med EPOS 2.0 och 1.0. 144
Tabell 39. Prestation med avseende på ordklassindelning av okända ord i EPOS 2.0 och 1.0. 145
Tabell 40. Indelning av fel från disambiguering med EPOS 2 och 1. (5083 resp. 325 st.) 150
Tabell 41. De 7 vanligaste regeltyperna vid disambigueringsfel i EPOS. 154
Tabell 42. De 7 vanligaste regeltyperna vid korrekta analyser i EPOS. 154
Tabell 43. Indelning av fel från analys av okända ord med EPOS 2 och 1. (929 resp. 24 st.) 156
Tabell 44. Fördelning av taggningfel mellan manuell och automatisk uppmärkning, Källgren (1998). 160
Tabell 45. De 10 mest frekvent feltaggade funktionsorden, Källgren (1996). 160
Tabell 46. De 10 mest frekvent feltaggade orden från utvärderingen av EPOS. 160
Tabell 47. Prestation i form av procent av fel i POS-taggning, McEnery och Wilson (1997:126). 164
Tabell 48. Studie av olika taggares prestation på samma text, Zavrel och Daelemans (1999) 165
Tabell 49. Studie av olika språk men samma taggningsmetod, Zavrel och Daelemans (1999) 165
Tabell 50. Prestation hos olika taggare med samma tagguppsättning och träningstext, Megyesi (2001) 166
Tabell 51. Prestation hos olika taggare med större tagguppsättning vid träning och taggning än vid utvärderingen, Megyesi (2001). 168
Tabell 52. Resultat från Tufis studie med tiered tagging, Tufis (2000) 169
Tabell 53. Elworthys studie av relation mellan tagguppsättning och korrekthet, Elworthy (1995). 169
Tabell 54. Exempel på ord som skiljer sig i taggning mellan SUC och SMDB. 174
Tabell 55. De senaste 10 ändringarna i mappningsmodulen. 176

Tabell 56. 'Höstas' disambiguerat före korrigering i mappning i EPOS. 177
Tabell 57. Disambigueringstyper med namn och innehåll i EPOS. 178
Tabell 58. 'Höstas' disambiguerat efter korrigering i mappning. 179
Tabell 59. Första exemplet på fel när kontexten är okänd i EPOS..... 179
Tabell 60. Andra exemplet på fel när kontexten är okänd i EPOS..... 180
Tabell 61. Tredje exemplet på fel när kontexten är okänd i EPOS. 180
Tabell 62. Exempel före kongruensanalys i EPOS..... 181
Tabell 63. Exempel efter kongruensanalys med EPOS..... 182
Tabell 64. Exempel före infinitivmärkesanalys med EPOS. 183
Tabell 65. Exempel efter infinitivanalys med EPOS..... 183
Tabell 66. Exempel före analys av 'som' i EPOS. 184
Tabell 67. Exempel efter analys av 'som' i EPOS. 185
Tabell 68. Exempel före analys med bigram i EPOS. 186
Tabell 69. Exempel efter analys med bigram i EPOS. 186
Tabell 70. Tabell med de 10 mest frekvent feltaggade orden i EPOS..... 186
Tabell 71. Taggуppsättning för EPOS 1.0. 213
Tabell 72. SMDB – taggуppsättning 217

Exempelförteckning

Exempel 1. Regelbeskrivning från TAGGIT, Hindle (1989)..... 35
Exempel 2. Ordet "kilometer" med olika numerusinformation i SUC 2.0. 47
Exempel 3. Ordet "finansplanen" ska vara NN UTR SIN DEF NOM i SUC 2.0. . 47
Exempel 4. Ordet "får" är nomen i stället för verb i SUC 2.0. 47
Exempel 5. Frasen "ta vara på" med tre olika taggar på 'vara' i SUC 2.0..... 47
Exempel 6. Frasen "en av" med tre olika taggar på 'en' i SUC 2.0..... 48
Exempel 7. Exempel på annoterad text i Press 65. 59
Exempel 8. Exempel på disambiguerad text i Press 65..... 60
Exempel 9. Exempel på taggning i SUC 2.0. 61
Exempel 10. Exempel på en flerordsenhet taggad med SMDB. 62
Exempel 11. Exempel på en flerordsenhet taggad i SUC. 62
Exempel 12. SMDB-annoterad text med homografer, okänt ord: "CSS2-missilerna". 111
Exempel 13. Disambiguerad text, med analyserat okänt ord, med EPOS. 111

Ekvationsförteckning

Ekvation 1. Sannolikhetsformel från CLAWS1 (Marshall, 1987 i van Halteren, 1999:12). 35
Ekvation 2. Formel för uträkning av precision, van Halteren (1999:82). 50
Ekvation 3. Formel för uträkning av recall, van Halteren (1999:82). 50
Ekvation 4. Formel för uträkning av F-measure, van Halteren (1999:82). 51
Ekvation 5. Formel för uträkning av Tag average per token, van Halteren (1999:82). 51
Ekvation 6. Formel för uträkning av korrekthet, van Halteren (1999:82)..... 52
Ekvation 7. Sannolikhetsformel för okända ord, med suffix och kontext i EPOS. 108
Ekvation 8. Sannolikhetsformel för okända ord, med suffix, utan kontext i EPOS.... 108
Ekvation 9. Bayes teorem för beräkning av sannolikhet för en företeelse givet en annan, Manning & Schütze (1999:43). 109
Ekvation 10. Formel för lexikalisk sannolikhet enligt beräkning med "maximum likelihood". (Brill 1999, i van Halteren 1999:210). 109
Ekvation 11. Förslag till formel för lexikalisk sannolikhet för okända ord med "one-count"-frekvens, (Brill 1999, i van Halteren 1999:210). 110
Ekvation 12. Uträkning av lexikalisk sannolikhet för okända ord, (Weischedel 1993, i van Halteren 1999:210). 110
Ekvation 13. Formel som anger balans mellan lexikalisk och kontextuell sannolikhet i CLAWS1, Marshall (1987:42-56)..... 115
Ekvation 14. Mönster för trigram av ordklasser, (El-Beze et al. i van Halteren 1999:271-272). 116
Ekvation 15. Mönster för trigram av ordklasser i EPOS..... 116
Ekvation 16. Mönster för trigram med ord och ordklasser i EPOS..... 116
Ekvation 17. Sannolikhet med relativ frekvens, Manning och Schütze (1999:197).... 121
Ekvation 18. Sannolikhetsformel för lexikalisk information med kontext i EPOS. 122
Ekvation 19. Sannolikhetsformel för lexikalisk-kontextuell information med kontext i EPOS. 122
Ekvation 20. Sannolikhetsformel för morfosyntaktisk information, med kontext i EPOS. 122
Ekvation 21. Sannolikhetsformel för ett token, utan kontext i EPOS. 122
Ekvation 22. Sannolikhetsformel för suffixinformation, utan kontext i EPOS..... 123
Ekvation 23. Formel för uträkning av homografifrekvens i ett lexikon i EPOS..... 132
Ekvation 24. Formel för uträkning av homografifrekvens i en text i EPOS. 132
Ekvation 25. Formel för uträkning av homografins relation till mått på recall i EPOS. 133
Ekvation 26. Uträkning av homografifrekvens i lexikonet i EPOS. 133
Ekvation 27. Uträkning av $H_{measure}$ baserat på H_{lex} i EPOS. 134
Ekvation 28. Uträkning av homografifrekvens i texten i EPOS..... 134
Ekvation 29. Uträkning av $H_{measure}$ baserat på H_{text} i EPOS. 134
Ekvation 30. Uträkning av medelvärde för H_{lex} och H_{text} i EPOS. 135
Ekvation 31. Uträkning av $H_{measure}$ baserat på H_{med} i EPOS..... 135
Ekvation 32. Sannolikhetsformel för n-gram med fyra ord i kontexten i EPOS..... 188
Ekvation 33. Sannolikhetsformel för bigram med ett ord i kontexten i EPOS. 185

Termförteckningar och sakregister

A
AI 30
algoritmer
  lärande algoritmer 33
angreppssätt
  dataimplementerade angreppssätt 25
  lingvistiska angreppssätt 25
artificiell intelligens (AI) 30
attribut
  obligatoriska attribut 42
  rekommenderade attribut 42

B
backtracking 116
baseline 128
batch learning systems 32
Bayes teorem (BT) 108
belagd homografi 22
bigrams 179
BT 108

C
Case-Based Learning 33, 103
Categorial errors 53
CES 59
CG 29
CGI 123
Common Gateway Interface (CGI) 123
compounds 14
connectionism 34
Constraint Grammar (CG) 29
Corpus Encoding Standard (CES) 59
Corpus Encoding Standard för XML (XCES) 59

D
Database Interface (dbi) 123
datorstyrda lokala regler 35
dbi 123
disambiguering 98
diskurs 24

E
enclitic forms 14
Ett Utvärderingsverktyg För Ordklasstaggning och gRammatisk parsning och felanalyS (EUFOROS) 139
EUFOROS 139

F
facit 63
fallbaserade taggers 36
Feature errors 53
feltyper
kategorifel 53
morfologiska fel 53
flerordsenheter 13, 27
F-measure 51
se även utvärderingsmått 51

G
grafem 16
graford
heterografa graford 16
homografa graford 16
guldstandard 63

H
Hidden Markov Models (HMM) 5, 36
$H_{measure}$ 132, 136
se även utvärderingsmått 132, 136
HMM 5
homografer 18
externa homografer 16
interna homografer 17
homografi
belagd homografi 22
extern homografi 18
interlingval homografi 17
intern homografi 18
partiell homografi 17
potentiell homografi 18
potentiella homografin 21
resterande homografi 21
homonymi 18

I
IE 168
inducerat inlärningssystem 30
inducerat inlärningssystem i EPOS 104
Information Extraction (IE) 168
informationsutvinning 168
inlärningstekniker
giriga inlärningstekniker 34
lata inlärningstekniker 34

K
Kappa-måttet 136

se även utvärderingsmått 136
klassbeteckning 17
korpus
källkorpus 64
målkorpus 64
träningskorpus 63
korrekthet 51
se även utvärderingsmått 51
kriterier
externa kriterier 44
interna kriterier 44
kriterier och riktlinjer 43

L
Language Engineering (LE)-Språkteknologi 13
LE 13
lemma 16, 18
lexikaliserad ordförbindelse 13
logogram 16

M
Machine Learning (ML) 30, 63
supervised learning 29
unsupervised learning 29
Maximum Entropy (ME) 5, 37, 161, 165
Maximum Likelihood 109
MBL 36, 165
ME 5, 37, 161, 165
Memory Based Learning (MBL) 36, 165
Mergers 14
miniminivå 128
ML 30, 63
morfologisk analys 97
morfosyntaktiska enheter 13
multiwords 13
se även word 13

N
Namned Entity Identification (NE) 52
Natural Language Processing (NLP) 24
NE 52
Neural Networks 34
neurala nätverk 36
N-gram 35, 114, 115
bigrams 115
trigrams 115
NLP 24

O

ord
- alfabetiska ord 16
- graford 16
- hybrida ord 16
- junkturella 16
- lexikonord 13
- numeriska 16
- ortografiska ord 13
- skenord 14
- textord 13

ordkropp 17

P

parsning 24
phantom words 14
- se även word 14

polysemi 18
potentiell homografi 18
precision 50
- se även utvärderingsmått 50

proclitic forms 14

R

recall 50
- se även utvärderingsmått 50

regelinduktion 103
relativ frekvens 109
resterande homografi 21
Rule and Decision Tree Induction 34

S

sammansättningar 14
SGML 59
smoothing 173
speciella extensioner 42
Standard Generalized Markup Language (SGML) 59
sublemma 17
supervised learning 29
syntaktiska idiom 27

T

Table Look-Up 33
Tag average per token 51
- se även utvärderingsmått 51

taggning
- tiered tagging 168

tagguppsättning 41

TBL 37, 40
tiered tagging 168
TnT 165
token 13, 22
tokenisering 97
Transformation Based Learning (TBL) 37, 40, 165
trigram 104, 114
  kontextuella trigram 120
  lexikaliska trigram 112, 120
  lexikalisk-kontextuella trigram 112, 120
  suffixbaserade trigram 120
  tokenbaserade trigram 120
Trigrams'n'Tags (TnT) 165

U
unsupervised learning 29
uppmärkning
  semantisk uppmärkning 24
  syntaktisk uppmärkning 24
utvärderingsmått
  F-measure 51
  $H_{measure}$ 132, 136
  Kappa-måttet 136
  korrekthet 51
  precision 50
  recall 50
  Tag average per token 51

W
word
  multiwords 13
  phantom words 14
  word token 13
  word types 13
World Wide Web (WWW) 123
WWW 123

X
XCES 59

Källförteckning över lexikaliska och textuella material och datorprogram

A
A probabilistic tagger for Swedish using the SUC tagset 39
A Robust and Modularized Lemmatizer/Tagger for Swedish 40
AVENTINUS 40

B
Baum-Welch-algoritmen 116
BETA 38
Brill tagger 40
Brown Corpus 25

C
CLAWS1 ..... 25, 35
Constituent Likelihood Automatic Word-Tagging System (CLAWS) 30

E
EAGLES 40, 42
En stokastisk ordklasstagger för svenska 39
EngCG 29
EPOS
    EPOS 1.0 56
    EPOS v.1.0 tagset 67
    EPOS-En Probabilistisk Ordklasstagger för Svenska 2, 96
Ett Utvärderingsverktyg För Ordklasstaggning och gRammatisk parsning och felanalyS (EUFOROS) 139
EUFOROS 139

F
Forward-Backward algoritm 116

G
GATE 40
GRANSKA 40

H
HMM-tagger 39

L
Lexikon Orienterad Parser för Svenska (LPS) 38
LexWare 40
LOB Corpus 25
lookup_saol 72
LPS 38

M
map_tagset 72

Ment Model 39
MorP 38
Morph 38
Morphological Parser (MorP) 38
Morphological Tagging Based Entirely on Bayesian Inference 39
µ-TBL 40
MULTEXT 40
MULTEXT-ISSCO bigram tagger 168

N
NEGRA-korpus 162
NFO 46, 57, 163
Nusvensk Frekvensordbok (NFO) 15

P
Press 65 ..... 57, 58

R
Rule-Based Tagging in Språkbanken 40

S
SAG 13, 15, 46
SAOL 46, 55
SAOL11 56
SAOL12 56
SAOL-Svenska Akademiens Ordlista v. 12 4
SAOL11 40
SCOL-parser 124
SMDB 46, 55
SMDB:s tagset 66
SMDB-Svensk Morfologisk Databas 2
Språkbanken 40
SUC..... ..... 46, 60
SUC:s tagset 64
SUC-Stockholm-Umeå Corpus 2
SWETWOL 38, 46, 60

T
TAGGIT 29, 35
The Xerox Part-of-Speech Tagger (XPOST) 39
TnT..... 165
TnT-tagger 162
Tokenize 105
Trigrams'n'Tags (TnT) 165

U
UCP..... 38
Uppsala Chart Processor (UCP) 38

V
Viterbi-algoritmen 116
Volsunga 115
VOLSUNGA-algoritmen 39
X
XPOST.. 39

Författarindex

A

Abney 124
Allén 14, 66, 190
Allén et al. 57
Allwood 57
Andersson 205

B

Barnbrook 23
Berg 106, 175
Berg et al. 66, 105
Berg och Cederholm 66
Blåberg 39
Brants 161, 168, 169, 190
Brill 37, 104, 109, 205, 206
Brill och Marcus 30
Brodda 38, 206
Bromley 38

C

Carlberger 40
Carletta 136
Chanod och Tapanainen 29
Chen 206
Church 115, 206
Cutting 36, 39, 55
Cutting et al. 30

D

Daelemans 30, 36
De Rose 25, 39, 115
Dura 40

E

Ejerhed 38
Ejerhed et al. 60
El-Beze 116
Elworthy 39, 44, 53, 169

F

Frakes och Baeza-Yates 48
Francis och Kucera 25

G

Garside 13
Garside et al. 30, 43
Gilberg och Armstrong 168

Greene och Rubin 25
Grefenstette 105
Grossman och Frieder 48

H
Halliday 28
Hassler-Göransson 57
Hindle 35

J
Johansson Kokkinakis 40

K
Källgren 38, 53, 60, 159, 189
Kann 40
Karlsson 38, 60, 163, 190
Kokkinakis 40
Kokkinakis et. al. 106

L
Lager 40
Leech 13, 14, 43, 208
Leech och Smith 23
Ljung/Ohlander 14

M
Manning och Schütze 36, 48, 109
Marshall 35, 115
McEnery och Wilson 26, 97, 164
Megyesi 40, 69, 161, 163, 165, 167, 190

N
Nagata 115

O
Oflazer och Kuruöz 29

P
Prütz 20

R
Ratnaparkhi 37

S
Sågvall 38
Samuelsson 39, 106
Schiller och Karttunen 106
Schmid 36
Schütze 36
Shannon och Weaver 28

Siegel och Castellan 136
Skut et al. 162

T
Tufis 69, 168

V
van Guilder 29
van Halteren 49
van Rijsbergen 48
Viterbi 116
Voutilainen 25, 29, 115, 164

W
Weischedel et al. 110, 115

Z
Zavrel och Daelemans 36, 164

Å
Åström 39, 115

1 Inledning

1.1 Bakgrund

Vid språkvetenskapligt arbete och bearbetning av texter av olika slag är morfosyntaktisk uppmärkning (ordklasstaggning) oftast ett av de första stegen. Ordklasstaggning kan utföras manuellt men också maskinellt vilket är vanligare. Ungefär 50% av alla ord i en text är homografa. Det innebär att de kan ha minst två ordklasstaggar beroende på vilken kontext de förekommer i. Hur mycket *homografi* en text består av beror också på hur specifik ordklasstaggningen görs. Ibland uppstår homografi inom en ordklass om tillräckligt specifik morfologisk information anges. Homografi kan synliggöras genom att ett lexikon, med fler ordklasstaggar än en för vissa ord, används för att göra en första uppmärkning av ord i en text. Nästa steg i ordklasstaggningen är disambigueringen, dvs. att avgöra i vilket sammanhang en viss ordklasstagg är rätt. Disambigueringen försvåras av att många ord i en text tolkas som homografa. Även om homografin alltid finns potentiellt i en text beror den dessutom på vilket lexikon som används. Ett begränsat lexikon som inte innehåller mycket homografi underlättar disambigueringsprocessen. På motsvarande sätt försvåras disambigueringen då ett mer fullständigt lexikon med mycket homografi används. Ur prestationssynpunkt är ett begränsat lexikon att föredra men ur korrekthets- och fullständighetssynpunkt är det bättre med ett mer fullständigt lexikon.

Förutom att homografi i ett lexikon och en text försvårar ordklasstaggning, beror utgången av hur korrekt ordklasstaggningen blir också på vilken *datateknisk metod* som används, förutsatt att processen utförs maskinellt. Det finns flera metoder och de mest framgångsrika är baserade på statistiska tekniker.

De ovan nämnda faktorerna påverkar resultatet av ordklasstaggning på ett direkt sätt då faktiska förändringar i definition av homografi och datateknisk metod ger direkta resultat. Det finns dock andra faktorer som på ett mer indirekt sätt anger hur resultatet uppfattas. Det kan vara hur ordklasstaggarna, som används i ordklasstaggningen definieras, grupperas och sätts samman. Detta kan sammanfattas under beteckningen taggupp-sättning. Även vilken utvärderingsmetod som används vid utvärdering påverkar uppfattningen av resultatet.

En *taggupp-sättning* kan beskriva olika specifika ordklasstaggar. Ju mer specifik information som anges i taggarna, desto fler homografer bildas i lexikonet och texten. Därmed blir också disambigueringsuppgiften

svårare. Beroende på det språk i vilket en text är skriven, finns också olika mycket morfologisk böjningsinformation som eventuellt kan beskrivas i en tagguppsättning. Jämför här två språk som engelska, där föga böjningsmorfologi förekommer, med ungerska som innehåller mycket böjningsmorfologi. En mindre och mer begränsad tagguppsättning kan därmed användas för engelska medan en mer komplex och utförlig tagguppsättning krävs för ungerska.

Vilken *utvärderingsmetod* som används påverkar också resultatet av ordklasstaggningen på ett indirekt sätt. Indirekt eftersom själva resultatet i sig inte påverkas utan snarare uppfattningen av det. Det finns olika sätt att mäta hur korrekt en text är ordklasstaggad. Vissa mått är mer generella och andra mer specifika. De generella måtten ger en övergripande bild av korrektheten och är de som används mest. De tar inte hänsyn till olika påverkande faktorer. Mer specifika metoder bör utvecklas i syfte att just ta hänsyn till olika faktorer som påverkar svårigheten att disambiguera.

1.2 Syfte

Avhandlingsarbetet syftar till att utforska faktorer och problem som påverkar prestation i automatisk morfosyntaktisk uppmärkning med praktiska exempel från projektet, att tagga texter i Språkbanken med de tre givna resurserna Svensk Morfologisk Databas (SMDB), En Probabilistisk Ordklasstaggare för Svenska (EPOS) och Stockholm-Umeå Corpus (SUC).

Den problemställning som avhandlingen behandlar är att bestämma i vilken grad det går att korrekt morfosyntaktiskt uppmärka texter med givna resurser. Frågan besvaras genom att olika faktorer beaktas som både direkt och indirekt kan anses påverka prestationen. Dessa är lexikaliskt och textuellt utgångsmaterial, disambigueringsmetod, tagguppsättningsinnehåll och -komposition samt vilken typ av utvärderingsmetod som används vid bedömning av prestationen. Förutom faktorer som påverkar prestationen diskuteras också vilken som är den mest lönsamma ansatsen när det gäller att förbättra prestationen. Tre möjliga vägar undersöks. De är förbättring av träningsmaterial, statistiska disambigueringsmetoder och försök med vissa punktinsatser inom givna problemområden. Avslutningsvis diskuteras hur prestation av korrekt taggade texter bedöms. Det gäller vilka mått som normalt används och hur de speglar viktiga påverkande faktorer. Möjliga faktorer kan vara storlek och fördelning av tagguppsättning, kvalitet och storlek på träningsmaterial och homografifrekvens i lexikon och text som ska analyseras.

1.3 Tillvägagångssätt i avhandlingsarbetet

Det går att med hög korrekthet (94,04% med EPOS v. 2.0 2002, respektive 96,5% med EPOS v. 1.0 1999) morfosyntaktiskt uppmärka texter med de resurser och de metoder som används i det arbete inom Språkbanken som används för exemplifiering i avhandlingen. Eftersom resultatet av ordklasstaggning är beroende av tillgängliga resurser och påverkas av diverse faktorer måste dessa redovisas. Nedan presenteras några faktorer som direkt eller indirekt påverkar prestationen. Avhandlingen omfattar en utförlig felanalys och utreder även möjligheter till förbättring av prestation och diskuterar till vilket pris förbättringar kan utföras och hur prestation normalt bedöms i ordklasstaggade texter. Dessutom föreslås ett nytt sätt att bedöma homografifrekvens i lexikon och text i förhållande till ytliga prestationsmått.

Delar av avhandlingsarbetet har utförts inom ett taggningsprojekt i Språkbanken. Detta har resulterat i taggningsverktyget (EPOS) vilket har använts för att, så utförligt som möjligt, morfosyntaktiskt tagga alla press- och romanmaterial i Språkbanken. Det kommer också att vara tillgängligt för taggning av andra textmaterial. Detta nyutvecklade verktyg redovisas och jämförs med liknande taggningsverktyg, avseende metod och teknik. Dessutom diskuteras och exemplifieras en mängd problemställningar av allmänt intresse som uppkommit i arbetet med att utveckla verktyget. Ofta handlar det om problem som inte har någon given lösning.

1.3.1 Faktorer som påverkar prestationen i taggning

Det är uppenbart att det finns faktorer som i olika grad påverkar prestationen i ordklasstaggning och därför är det viktigt att dessa faktorers inverkan analyseras och att det förs en diskussion om i vilken utsträckning de påverkar möjligheterna att tagga korrekt. I denna studie kommer fem möjliga direkta och indirekta faktorer att tas upp. De direkta är: träningsmaterial, lexikaliskt material och disambigueringsmetod i taggningsverktyg. De indirekta är: tagguppsättning och utvärderingsmetod. De direkta faktorerna har ett orsakssamband med prestationen medan de indirekta snarare påverkar tolkningen av prestationen.

Kvalitet och storlek på *träningsmaterial* (SUC) som har använts i EPOS har haft stor betydelse för prestationen i taggningen. Kvaliteten har försämrats något då det i mappningen har uppstått problem, som medför att viss text inte kan användas vid regelformulering, vilket i sin tur innebär att vissa ord och uttryck inte representeras i disambigueringsverktyget. Det kan i vissa fall få stor betydelse då disambiguering av vissa vanliga ord

inte kan utföras med någon ledning av kontext utan slumpvis får en av grafordets möjliga taggar. När det gäller storlek på träningsmaterial har allmänt rekommenderad fördelning tillämpats. Detta innebär att 90% (ca 900.000 ord) av korpusen används för träning av taggningsverktyget och 10% (ca 100.000 ord) används för utvärdering.

Med *lexikaliskt material* avses det lexikon som används som utgångspunkt för disambiguering. Lexikonet som används är SMDB. Eftersom det är en maskinläsbar fullformsversion av Svenska Akademiens Ordlista v. 12 (SAOL) kan det anses vara det bästa och mest korrekta tillgängliga lexikaliska material, som finns att tillgå vid datamaskinell behandling av skriven svenska. Dess höga homografifrekvens bör egentligen inte vara en nackdel, då det handlar om att så utförligt och så verklighetstroget som möjligt beskriva ett graford i ett lexikon och i en text. De flesta ordklasstaggningssystem bedöms enbart efter korrekthet och då spelar homografin en stor roll för prestationen. Ju större homografin är desto fler taggar existerar vid disambiguering av homografa graford. För att på ett rättvisande sätt mäta prestation med homografi som en påverkande faktor föreslås i avhandlingen ett nytt sätt att bedöma prestation i utvärderingen av EPOS.

Taggningsverktyg som använts i arbetet är EPOS och *disambigueringsmetod* som används är en stokastisk analys med trigram. Metoden som har använts i detta arbete utgår ifrån de verktyg som fanns då denna studie påbörjades för fyra år sedan. Vid jämförelse av resultatet med användning av de taggningsverktyg som då var tillgängliga är resultatet godtagbart med tanke på att man kan anta att en s.k. minimigräns (*baseline*) för svenska ligger på ca 90%. Den anger en minimigräns för prestation oavsett metod. Minimigränsen kan uppnås genom att den mest frekventa taggen för ett homograft graford tilldelas ordet, oberoende av kontext. Dessutom är prestationen med verktyg som finns tillgängliga idag inte jämförbar med de resultat som kan uppvisas för engelska, framför allt därför att svenskan har en mer komplex böjningsmorfologi än engelskan (Megyesi 2001). Resultaten i denna avhandling visar att man endast kan förlita sig på statistisk analys till en viss gräns. Det är därför mer lönsamt att satsa på punktinsatser, när den statistiska analysen inte räcker till.

Den *tagguppsättning* som används i EPOS har skapats i utvecklingen av SMDB. Den bygger på den standardiserade rekommendationen av EAGLES. Den har i tidigare arbeten, fast då i större format, använts i taggningssystem som varit föregångare till EPOS. Storleken på en tagguppsättning har betydelse vid bedömning av prestation vid disambiguering. Ju större tagguppsättning desto svårare blir disambiguering av homografer därför att det då finns fler taggar att välja mellan. Enbart

storleken på taggуppsättning säger inget om fördelningen mellan olika ordklasser eller vilken morfologisk information som finns representerad i dem. Då olika morfologiska särdrag är olika svåra att analysera, bör taggningssystem med olika taggуppsättning inte jämföras. Sådana svårigheter är också ofta språkberoende. Eftersom SUC med en annan taggуppsättning har använts vid träning av EPOS blir jämförelser mellan SUC:s och SMDB:s taggуppsättningar dessvärre oundvikliga. Problem uppstår också i översättningen från SUC:s taggуppsättning till SMDB:s då alla taggar i SUC inte har en direkt motsvarighet i SMDB. I vissa fall motsvaras en tagg i SUC av flera taggar i SMDB, ett problem som diskuteras i metodavsnittet. Den taggуppsättning som används i SMDB och EPOS består av 133 taggar och den i SUC består av 155 taggar. Skillnaden i storlek beror på att det i SUC finns sammansättnings- och förkortningsdistinktioner, vilka inte finns i SMDB. I SMDB finns däremot fler specifika distinktioner inom adjektiv och nomen vad gäller genus och numerus, där det i SUC används mer generella övergripande taggar.

Utvärderingsmetod är en indirekt faktor som inte påverkar prestationen i sig, men som givetvis har betydelse för värderingen av prestation. Beroende på hur utvärderingen utförs och beroende på vad som utvärderas kan olika resultat uppnås. Definitioner och frågor som rör utvärdering diskuteras i prestationsdefinitionen, forskningsöversikten och vidare i ett separat avsnitt om utvärdering.

1.3.2 Förbättring av prestation

I de fall då prestation av ordklasstaggning betyder grad av korrekthet finns ett antal alternativ att tillgå; de som diskuteras i denna studie är förbättring av träningsmaterial, förbättring av statistiska metoder och punktinsatser. Den viktigaste frågan är vad som ger störst utdelning med lägst arbetsinsats.

Förbättring av *träningsmaterial* innebär att de problem som uppstår i mappningen mellan SUC:s taggуppsättning och SMDB:s korrigeras för att göra träningskorpusen fullständig. Det medför att de regler som automatiskt skapas utifrån korpusen också blir fler. I ett exempel i undersökningsavsnittet visas att effekten av att korrigera endast ett problemfall kan öka korrektheten i disambiguering för ett enskilt ord med ca 75%.

Förbättring av *statistiska metoder* kan göras genom att använda andra metoder, exempelvis Hidden Markov Models (HMM), eller Maximum Entropy (ME) eller liknande, vilket möjligtvis kan förbättra den statistiska analysen marginellt. Eftersom homografifrekvensen i lexikonet är den

största anledningen till att skillnaden i prestationen mellan EPOS 1.0 och EPOS 2.0 är så stor och det faktum att de bygger på samma metod, dras slutsatsen i denna studie att förbättringsansatser i stället bör inriktas på punkter där brister i statistisk analys uppvisas.

Punktinsatser kan fungera som ett komplement till den statistiska analysen. De kan användas för speciella analyser där statistiska metoder ger för osäkert underlag för att dra slutsatser. Insatser kan utföras både genom manuell utformning av disambigueringsregler och med statistiska medel. Problemområden kan identifieras genom en detaljerad felanalys. En sådan beskrivs i utvärderingsavsnittet.

1.3.3 Bedömning av prestation

Prestation är avgörande för hur metoder och verktyg uppfattas. Bedömning av prestation görs för att mäta hur väl och hur korrekt t.ex. ordklasstaggning utförs.

För att mäta prestation i ordklasstaggning används ofta relativt ytliga bedömningsmått. Exempel på detta är precision och recall (korrekthet). Måtten är ytliga i den bemärkelsen att de mäter mäter prestation över en hel text. Måtten ger ingen information om fördelning av prestation mellan ordklasser. De är dock enkla att använda och mycket vanliga. De kan däremot ge missvisande information om de används för att jämföra olika taggningsverktygs prestation då verktygen inte använder identiska lexikon, träningstexter och tagguppsättningar. Därför bör inte för stor vikt fästas vid dem, utan de bör tolkas som generella mått för att få en ungefärlig uppfattning om prestation i förhållande till vissa utgångsresurser. De utgångsresurser som bör tas hänsyn till är lexikonets enhetsförråd och informationsstruktur, träningstextens kvalitet och storlek och tagguppsättningens storlek och innehåll.

Vad som ska ingå i en prestationsbedömning är dels beroende på vad syftet har varit med taggningen och dels beroende på vad den uppmärkta texten ska användas till. Hur bedömningen ska utföras beror också på vad som ska visas i utvärderingsresultatet. Det vanligaste förfaringssättet är att endast generella och övergripande resultat redovisas. Om det å andra sidan är viktigt att framhäva andra faktorer, t.ex. storlek på tagguppsättning, träningskorpus eller homografifrekvens, måste sådan information också avspeglas i utvärderingsresultaten. Då syftet är att förbättra taggningsresultat måste en ingående, mer detaljerad, analys utföras så att eventuella fel kan identifieras och kategoriseras.

En central frågeställning som följer av ovanstående resonemang är hur man i bedömningen speglar faktorer som bevisligen försämrar de ytliga prestationsmått som normalt används. För att det ska vara möjligt att bedöma ovan nämnda faktorer i förhållande till reguljär prestation krävs andra metoder än de som används idag. Idealt skulle olika bedömningsmått därför användas för olika specifika faktorer och normal prestation skulle påverkas av dessa faktorer på olika sätt. Taggning med en större tagguppsättning vilket beskriver mer detaljerad morfologisk information än vanligt skulle höja värdet på prestationen ytterligare jämfört med taggning med en mindre tagguppsättning, vilken är mindre detaljerad. Ett taggningsverktyg som tränats på en liten korpus skulle få högre prestationsresultat i förhållande till om samma verktyg hade tränats på en större korpus. På samma sätt skulle prestationen bedömas högre vid användning av lexikon och text som innehåller en hög grad av homografi till skillnad från ett prestationsmått med lägre homografifrekvens då prestationen istället minskas. I denna avhandling föreslås ett nytt sätt att väga in mått på homografi i reguljära, ytliga utvärderingsmått.

1.3.4 Taggningsverktyget EPOS

Taggningsverktyget EPOS v. 2.0 har utvecklats i ett projekt för att morfosyntaktiskt uppmärka texter i Språkbanken. Det är en regelbaserad statistisk taggare som använder SMDB som lexikalisk resurs och SUC-korpusen som textuellt utgångsmaterial för träning och uppbyggnad. Eftersom EPOS och SMDB använder en annan tagguppsättning än den som finns i SUC, var mappning av tagguppsättningar nödvändig. I mappningen uppstod diverse problem som beror på olikheter vid tolkning av ord som gjorts vid utveckling av SUC och SMDB. Det beror också på olika undergrupperingar inom vissa ordklasser i de olika tagguppsättningarna. På grund av dessa problem är utgångsmaterialet fortfarande ofullständigt på de punkter där mappning inte var möjlig. Ungefär 95% av alla problemfall kommer att kunna åtgärdas. Det gäller dock inte ord där olika tolkningar gör att överbryggande inte är möjligt.

Träningsmetoder som använts vid uppbyggnad av disambigueringsregler i EPOS är dels en 'machine learning'-teknik, *Case-Based Learning*, därför att både tekniker som beskriver likhet eller närmaste motsvarighet till ett sökt regelmönster tillämpas, dels *regelinduktion*, eftersom den teknik som tillämpas generaliserar när sökt regelmönster saknas.

Metoden som används i disambigueringen är en enkel statistisk regelbaserad metod som bygger på trigram. Den uppvisar resultat som ligger betydligt högre än miniminivå. Som visas i utvärderingen har metoden inte

avgörande betydelse för prestationen. I en utvidgning av arbetet kommer det dock att finnas möjligheter att undersöka prestation med andra metoder.

Punktinsatser har utförts och redovisats i några problemområden. Ytterligare förbättringar kommer att göras på de punkter där brister i den statistiska analysen har uppvisats.

1.4 Avhandlingens uppläggning

Avhandlingen inleds med en forskningsöversikt i kapitel 2, i vilken termer som ord och ordklass definieras. Motivering till varför ordklasstaggning behövs och frågeställningar som uppstår vid ordklasstaggning diskuteras. En kort historisk bakgrundsbeskrivning ges till taggning och därefter en genomgång av ett taggningssystems beståndsdelar. Metoder och tekniker som används i dylika system beskrivs kortfattat. Slutligen sammanfattas taggnings- och morfologiska analyssystem utvecklade för svenska.

I kapitel 3 redogörs för de lexikaliska och de textuella material som använts i avhandlingen. De lexikaliska materialen är de lexikaliska databaser som använts och de textuella materialen är de korpusar som använts.

I kapitel 4 beskrivs vilka metoder som använts dels vid mappning av SUC:s tagguppsättning till SMDB:s och EPOS:s tagguppsättning, dels vid träning av EPOS, med utgångspunkt från SUC-korpusen. Slutligen redovisas också vilka metoder som används i EPOS vid disambiguering av homografer.

I utvärderingskapitlet, kapitel 5, definieras vanliga utvärderingsmått. Ett nytt förslag till mått för att väga in faktorn homografifrekvens och dess påverkan på prestationsmått ges. Därefter presenteras en utvärdering av prestation fördelat på ordklasser och en analys av fel indelat i kända och okända ord.

I kapitel 6 behandlas vissa faktorer som påverkar prestationen vid ordklasstaggning dels generellt och dels med de givna förutsättningar för avhandlingsarbetet.

En undersökning om möjliga förbättringar dels av mappningen, och dels av disambigueringen, redovisas i kapitel 7.

I kapitel 8 presenteras de slutsatser som kan dras om faktorer som påverkar svårigheten av ordklasstaggning, med utgångspunkt i det praktiska

arbete och de undersökningar som har utförts.

Slutligen anges litteraturreferenser.

Huvudtexten kompletteras med två appendix, enligt följande;

APPENDIX A) Tagguppsättning för EPOS 1.0

APPENDIX B) Tagguppsättning för SMDB/EPOS 2.0

Dessutom listas inledningsvis;

Innehållsförteckning

Figurförteckning

Tabellförteckning

Exempelförteckning

Ekvationsförteckning

Termförklaringar och sakregister

Källförteckning över lexikaliska och textuella material och datorprogram

Författarindex

2 Forskningsöversikt

Vid en diskussion om möjliga faktorer som påverkar prestation i ordklasstaggning är det nödvändigt att beskriva tidigare forskning anknuten till ämnet. Det görs i detta kapitel med en gruppering som följer den i avhandlingen tidigare gjorda och motiverade indelningen av direkta och indirekta faktorer. Eftersom översikten är indelad i två delar presenteras här kortfattat vad de olika delarna innehåller. Under direkta faktorer diskuteras först lexikaliskt och textuellt material och därefter disambigueringsmetod. I avsnittet om indirekta faktorer ges bakgrund till tagguppsättningar och utvärderingsmetoder. De definitioner och beskrivningar som ges gäller framförallt svenska språket. Det kan även finnas likheter med andra, främst nordiska, språk, men då denna avhandling endast behandlar svenska språket och eftersom författaren har valt att lägga ner mycket arbete på och beskrivning av den empiriska/praktiska delen av avhandlingen kan det på grund av tids- och utrymmesbrist inte dras några språktypologiska paralleller till andra språk.

2.1 Direkta faktorer

Avsnittet om lexikaliskt material diskuterar lexikoninnehåll och ordbegreppet under rubrikerna Enhetsförråd och Informationsstruktur. I avsnittet om enhetsförrådet anges de allmänna begrepp som krävs för att beskriva innehållet i ett lexikon, t.ex. ord, 'token' och lexikonord. I avsnittet om det lexikaliska materialets informationsstruktur diskuteras i vilken form ord kan representeras, t.ex. med information om ordklass och lemma, eller på vilken nivå ordet beskrivs, t.ex. som grafem, graford eller som fras. Detta avsnitt bygger delvis på den tidigare diskussionen om det lexikaliska materialets enhetsförråd. I avsnittet om Informationsstruktur ingår dessutom en diskussion om huruvida graforden är homografa eller heterografa.

I avsnittet om det textuella materialet beskrivs kvalitet och innehåll i textkorpusar, både efter egen och andras erfarenhet av arbetet med SUC. När det gäller innehållet i en textkorpus redogörs för vad begreppet homografi kan betyda, dels före disambiguering och dels efter. Det hänvisas vidare till ett antal studier om korpusars storlek vilket också nämns i kapitel 6.

Den tredje av de direkta faktorerna är disambigueringsmetod. Avsnittet inleds med en diskussion om vad ordklasstaggning innebär. Ordklasstaggningen ges också en historisk översikt. Vidare redogörs för uppbyggnaden av ett generellt ordklasstaggningssystem följt av en sammanfattning av några svenska ansatser i ämnet. I avsnittet diskuteras också skillnaden

mellan datadrivna och lingvistiska metoder, där de datadrivna, vilka också är de vanligaste exemplifieras vidare med tekniken *machine learning*. En kortare genomgång görs också av olika frekvent använda metoder, samt en beskrivning av en ny teknik som använder kombinationer av flera metoder.

2.1.1 Lexikaliskt material

Av det som kan räknas till lexikaliskt material är det främst lexikonets uppbyggnad och innehåll och lexikonet som en systematisk samling av datamaskinellt lexikaliskt material som är relevant för denna avhandling. Därför är det viktigt som utgångspunkt för vidare diskussioner att redogöra för vad som menas med ett ord som begrepp. Det är inte heller självklart hur en ordklass skall definieras. Det finns därför anledning att diskutera olika förslag, dels från lingvistiskt, dels från språkteknologiskt håll, för att kunna välja lämpliga modeller för fortsatta diskussioner om det som behandlas i avhandlingen. Författaren har valt att följa de definitioner vilka är mest frekvent använda vid beskrivning av dessa termer. Nedan följer en mer detaljerad beskrivning av hur ord och ordklasser kan definieras, avsnitt 2.1.1.1. I avsnitt 2.1.1.2 anges från Nusvensk Frekvensordbok (NFO) hur ord kan indelas i olika beskrivningsnivåer samtidigt som en definition av homografi ges.

2.1.1.1 Enhetsförråd

I anknytning till hur ett lexikons enhetsförråd är uppbyggt kommer nedan att diskuteras vad ett ord är och hur olika typer av ord benämns. Vad en ordklass är och hur den kan definieras är också relevant i denna diskussion.

Textord, token och lexikonord

Vad ett ord är och hur det avgränsas är något av det som beskrivs nedan med några exempel från bl.a Ljung/Ohlander (1982), Andersson (1996) och Svenska akademiens grammatik (1999), hädanefter SAG, vilka kan anses som representativa för den lingvistiska traditionen. Ljung/Ohlander beskriver ord, dels som helheter i form av signaler och dels som uppdelbara i mindre enheter såsom ljud och bokstäver. Fonemen byggs samman till morfem, minsta betydelsebärande enheter i språket, vilka sedan kombineras till ord. De beskriver alla språkliga enheter som har betydelse som "språkliga symboler" eller "språkliga tecken". Kännetecknande för det språkliga tecknet är att det utgör en kombination mellan en viss ljud- eller bokstavssekvens och en viss betydelse. Vidare skiljs ordbildning genom avledning från ordböjning. Böjningsändelser används för

att markera hur individuella ord förhåller sig till olika grammatiska kategorier medan avledningsändelser kan användas för att bilda nya ord.

Andersson (1996:19) beskriver ord på följande sätt: "I skriven text fungerar det bra att definiera ordet som en enhet som avgränsas av mellanslag." Men han diskuterar också problemet med var man ska sätta mellanslag i en text. Detta anser han beror på vilka kombinationsregler man använder vid ordkonstruktion, vilket han refererar till som skillnaden mellan ordbildning och ordfogning.

SAG gör följande distinktioner i fråga om definitioner av ord: *Lexikonord* är orden i talarens ordförråd och består av rotmorfem eller av flera morfem som är fast sammanfogade till lexikonord genom sammansättning eller avledning, t.ex. "smal, långsmal, smalna". Lexikonorden är de grundstenar som grammatiken är uppbyggd på. *Lexikaliserad ordförbindelse* är ett mellanting mellan ett flermorfemigt lexikonord och en syntaktiskt konstruerad ordgrupp och har en enhetlig betydelse vilken inte utan vidare kan förutsägas utifrån betydelsen hos de enskilda ord som ingår i förbindelsen, t.ex. "ge upp". De lexikaliserade ordförbindelserna kan också indelas i två grupper, fixerade enheter i vilka de ingående orden är omöjliga att böja eller skilja åt med andra ord, t.ex. "över huvud taget/ överhuvudtaget". Den andra gruppen innehåller ej fixerade enheter i vilka de ingående delarna kan splittras upp och eventuellt böjas, t.ex. "ger inte upp", "ångrade han sig". Den första gruppen kan jämföras med flerordsenheter.

Inom LE (Language Engineering-Språkteknologi) brukar man vara mer inriktad på enheter som påträffas i faktisk text och det är därför naturligt att börja med att definiera *Word token*. Den används för att ange en teckensekvens, där en sekvens omges av mellanrum (space) och det kan jämföras med svenskans *textord*.

Leech (Garside et al. 1997:21-24), definierar *ortografiska ord* som avgränsade enheter i skriven text vilka föregås och följs av mellanslag. Dessa skiljer sig från *morfosyntaktiska enheter* vilka han definierar som de textord (word tokens) vilka man identifierar vid grammatisk taggning. *Word token* definieras i sammanhanget som "...each instance of a word in a text counts as one token...", till skillnad från *word types* "...each word as listed in a dictionary, of which tokens are instances...". Leech påstår vidare att det finns tre avvikelser till ortografiska ords 1:1-förhållande till morfosyntaktiska textord[1]; 1) *Flerordsenheter* (multiwords): "... more than one orthographic word corresponds to one morphosyntactic word.". Han ger vidare exempel på detta med "in spite of", vilket kan tolkas som tre ortografiska ord, men de kan med fördel taggas som en enhet med ordklassen

preposition. 2) *Mergers*: " ... one orthographic word corresponds to more than one morphosyntactic word." Exempel på dessa kan vara *proclitic forms* såsom " t' " i franskans "Je t'aime" eller *enclitic forms* såsom " n't " i engelskans "hasn't". Dessa ord skrivs ofta ihop till en ortografisk enhet. 3) *Sammansättningar* (compounds): "... depending on the analysis, **one** or **more than one** orthographic word corresponds to **one** or **more than one** morphosyntactic word." Vidare definierar han sammansättningar på följande sätt: "...a compound may be defined as a word which has other words as its components." Var gränsen går mellan vad som kan anses vara en sammansättning och vad som kan ses som en lexikalisk enhet i ett lexikon kan diskuteras. En lexikalisk enhet, t.ex. "rainbow" vilket i och för sig kan delas upp i två delar, "rain" och "bow", men ändå står som ett enkelt nomen i ett lexikon, kan jämföras med "word class" (även "word-class"), vilket Leech anser vara en sammansättning då de båda ingående komponenterna "word" och "class" är nomen. Han menar att det ibland i engelskan är vanligt med bindestreck då det blir tydligare att det rör sig om en sammansättning. Leech ger ett annat exempel på så kallade "*skenord*" (phantom words) vilka kan uppstå i ortografisk analys, t.ex. "York-San" från ordsekvensen "New York-San Francisco flights", vilket kan uppstå då ortografiska ord ska motsvaras av morfosyntaktiska ord. Liknande problem har även framkommit i arbetet med Nusvensk Frekvensordbok (Allén 1970).

I detta avhandlingsarbete används de uppdelningar som har angivits av Leech (1997) och Allén (1970). En teknisk lösning till problemet är att identifiera flerordsenheter, eventuellt med bindestreck, för att beteckna en större sammansättning som då skulle bli "New York-San Francisco".

När det gäller formlära och ordklassindelning följer författaren även där de definitioner vilka är mest frekvent använda vid beskrivning av ordklasser; Nedan ges exempel från Ljung/Ohlander (1982) och SAG (1999).

Ordklassindelning

Ljung/Ohlander (1982:33) beskriver formläran och dess indelning av ord i olika ordklasser och man påpekar att "varje vetenskap behöver ett begreppssystem för att beskriva den aspekt av 'verkligheten' som den är intresserad av". De anser också att "gemensamt för alla språk är att de tycks ha bestämda regler beträffande de olika ordklasserna". Ordklasserna omfattar delvis ord med liknande innehållsliga egenskaper, men viktigast

1. Diskussionen kan ses som relativt språkoberoende då han ger exempel på andra språk än engelska, dock inte för språk utan ortografisk orddelning.

för att bestämma ordklass är ordens ändelser och ordföljden.

I SAG (Del 2, §4) beskrivs ordklasser och kriterier för ordklassindelning: "orden indelas i klasser efter hur de berörs av grammatikens regler för böjning av ord samt för konstruktion av fraser och satser och efter hur de systematiskt bidrar till de syntaktiska konstruktionernas betydelse." I SAG delas ordklasserna in i tre olika grupper, beroende på hur rik böjning ordklasserna har; "Ord som tillhör ordklasserna substantiv, adjektiv och verb har en rik böjning och avtecknar sig tydligt mot de oböjliga ordklasserna prepositioner, konjunktioner, subjunktioner, infinitivmärke och interjektioner. De övriga ordklasserna egennamn, pronomen, räkneord, particip och adverb intar i fråga om böjning en mellanställning."

Ord som ligger utanför det egentliga grammatiska systemen men ändå ofta uppmärks i kategorier är utländska ord och olika typer av förkortningar eller sammansättningar. I SUC:s tagguppsättning används alla tre men i SMDB används endast de två första.

Vidare diskussion om ordklasser finns i nästa avsnitt om beskrivningsnivåer från NFO, avsnitt 2.1.1.2.

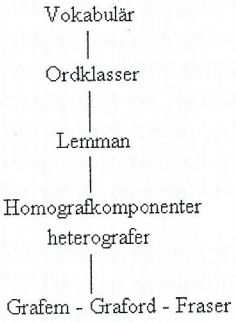
2.1.1.2 Informationsstruktur

Ett lexikons informationsstruktur bestämmer vilka typer av information som lagras. En lämplig bakgrund för en diskussion om informationstyper utgörs av beskrivningsnivåerna i NFO, vilka sammanfattas nedan. Sammanfattningen knyter delvis an till de beskrivningar som tidigare gjorts av ordbegrepp och ordklassindelning. Beskrivningsnivåerna är dessutom en lämplig introduktion till homografi- och homonymibegreppet som kan anses vara centralt i denna avhandling. Här talas dock främst om homografi eftersom det rör sig om skrivet material. Homografi i lexikonet redogörs för vidare i ett separat avsnitt. Här kan även nämnas att homografi återkommer i diskussionen om det textuella materialet, avsnitt 2.1.2, då homografin behandlas ur ett annat perspektiv.

Beskrivningsnivåer av ord från Nusvensk Frekvensordbok (NFO)

För att knyta samman definitionen av ord, ordklasser, ordklasstaggning och homografseparering ges nedanstående beskrivning hämtad från NFO, Allén et al. (1970). Anledningen till att just NFO används är att SMDB i mycket bygger på NFO-modellen. Vid uppbyggnaden av frekvensordböcker som NFO finns dessutom ett krav på att de måste vara formellt väldefinierade på ett sätt som är användbart inom språkteknologin.

I NFO används olika beskrivningsnivåer för att representera den grammatiska information som kan tillhöra ett ord. Ett antal beskrivningsnivåer definieras och framgår av figur 1.



Figur 1. Beskrivningsnivåer från NFO, Allén (1970).

Här beskrivs den grundläggande ordnivån, *grafordet*, vilket är det grafiska ordet. Det är en enhet som bestäms såsom ett segment begränsat av två på varandra följande spatier (mellanrum). Graforden är i sin tur uppbyggda av *grafem*. Grafemuppsättning kan vara en utgångspunkt för klassificering av ord. Ordtyper kan separeras i fyra olika grupper: de *alfabetiska* orden vilka är uppbyggda av alfabetiska grafem (bokstäver), de hybrida orden inkluderar alfabetiska, *numeriska* (siffror) och *junkturella* (skiljetecken) och *logogram* (%, & etc.).

Närmast högre nivå omfattar graford som kan vara *homografa* eller *heterografa*. Ett exempel på en homograf ordform är 'stack' vilket både kan vara ett nomen och ett verb. Heterografa ord är icke-homografa ord. Ett exempel på ett homograft ord är 'arm' som kan tillhöra ordklassen nomen och adjektiv.

Nästa nivå i beskrivningen är ett ords lemmatillhörighet. Ett *lemma* kan definieras på följande sätt: "ett lemma är en grupp ordformer inom en ordklass vilka kan hänföras till antingen en och samma flexionsserie eller flera i tal och/eller skrift konvergerande serier vars divergenser visar rent fakultativ (fri) variation. Oböjliga ord omfattar endast grundformen." Det första steget i homografseparering för identifiering av lemmatillhörighet kan vara att skilja de ord åt som inbördes har olika ordklasser (*externa* homografer), liknande exemplet med 'stack' tidigare. En mer omfattande homografseparering krävs då ord inom samma ordklass tillhör olika lemman. Ett exempel på detta är ordet 'sticka', vilket t.ex. kan indelas i

två enheter, men inom ordklassen verb, då de två enheterna har olika böjningsserier, 'sticka, stack' respektive 'sticka, stickade'. En annan form av homografi finns mellan olika böjningsformer inom samma lemma (*interna* homografer), se exempel 'stickat' i kolumn 6 i tabell 1. En tredje typ av homografi är den som kallas *partiell homografi*, dit proprier och icke-proprier hör. Exempel på detta är 'Karl' i början av en sats vilket kan vara både proprium och icke-proprium. En fjärde typ av homografi är den så kallade *interlingvala* homografin, vilken gäller svenska ord som sammanfaller med samma ordformer på andra språk, t. ex. 'and' (eng.), 'le' (fr.) och 'den' (ty.). I tabell 1 framgår hur homografer och heterografer och lemman fördelar sig över graforden 'stacken, stack, sticker, sticka, stickar, stickat'.

C	stack		sticka			sticka	sticka					
	nn -en		vb -ø			nn -n	vb -ad					
B	stacken	stack nn -en	stack vb -ø	sticker	sticka vb -ø	sticka nn -n	sticka vb -ad inf	sticka vb -ad imp	stickar	stickat vb -ad sum	stickat vb -ad ptp	
A	stacken	stack		sticker	sticka				stickar	stickat		
	1	2		3	4				5	6		

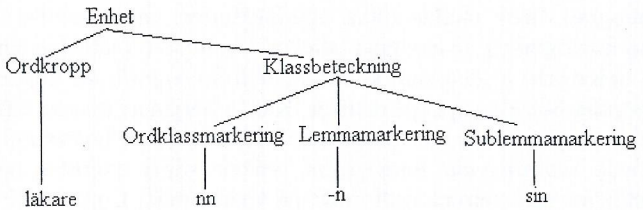
Tabell 1. Exempel på homografseparering och lemmatisering, Allén S. NFO (1970).

I tabell 1 kategoriseras graford, homografkomponenter och lemman på följande sätt.

A	grafordsnivå
B	homografkomponentnivå
C	lemmanivå
1, 3, 5	heterografi
2	extern homografi
4	extern och intern homografi
6	intern homografi

Tabell 2. Fördelning av graford, homografkomponenter och lemman från tabell 1.

Vid ordklasstaggning används i NFO begreppet *enhet*, där en enhet på grafordsnivå utgörs av en *ordkropp*. På homografkomponentnivå utgörs en enhet av en ordkropp och en *klassbeteckning*. Klassbeteckningen kan innehålla tre delar av information: *ordklass*, *lemma* och *sublemma*. I figur 2 illustreras denna indelning av begreppet enhet.



Figur 2. En homografkomponents uppbyggnad i NFO, Allén (1970).

I figur 2 består ordklassmarkering av en entydig förkortning, lemmamarkering av en böjningsangivelse och sublemmamarkering av flexionsform inom lemmat (vid intern homografi).

I en licenciatavhandling presenteras en studie av homografi i nusvenskan av Sture Berg (1970). I den redogörs också för en begreppsdefinition som rör *homonymi*. Där sammanfattas att begreppet homonym i olika källor anges som motsatsen till synonym, dvs. ett ord som uttalsmässigt låter och ser lika ut, men inte har samma betydelse som ett annat ord. *Homograf* definieras som "... ett ord som till sin skriftform är identiskt med ett annat ord men olika beträffande ljudform, ursprung och betydelse..." (s. 10). Vidare anges också från Malmberg (1962) *homonymi* som tillfällig ljudlikhet mellan ord, t.ex. 'skola' (som substantiv eller verb) och *polysemi* som mångtydighet, t.ex. 'blad' (i bok och på växter). Berg redovisar därefter sin egen ståndpunkt i frågan om definition av dessa termer. Den bygger på Alléns tidigare definitioner. Han skiljer på definitionen av homonymi där '/jä:rna/' (som i 'gärna' och 'hjärna') är ett exempel. Medan begreppet homografi kan förklaras med exemplet 'loge' (som används för förvaring av otröskad säd) och 'loge' (omklädningsrum för skådespelare). Vidare anger han följande definitioner; *extern homografi* "är den relation som råder mellan i skrift identiska former tillhörande olika lemman." (s. 27) *intern homografi* "är den relation som råder mellan i skrift identiska flexionsformer (grundformen inräknad) inom ett och samma lemma." (s. 27) Slutligen anges att *lemma* "är en grupp ordformer inom en ordklass vilka är hänförliga till antingen en och samma flexionsserie (i fråga om oböjliga ord endast omfattande grundformen) eller flera i tal och/eller skrift konvergerande serier vars divergenser visar rent fakultativ variation." (s. 26)

Den homografi som omtalats generellt i redovisningarna ovan kommer i fortsättningen av avhandlingstexten att benämnas som *potentiell homografi*, om inget annat anges. Men potentiell homografi kommer fortsätt-

ningsvis inte säga något om vilken typ av homografi det rör sig om utan om det faktum att ett givet lexikon förutsäger en viss möjlig grad av homografi för alla textanvändningar av lexikonets ord.

2.1.2 Textuellt material

Av det som kan räknas till textuellt material är det kvalitet, innehåll i korpusar och storlek på korpusar, samt hur de kan användas för ordklasstaggning och skapande av ordklasstaggningsverktyg, som är relevant för denna avhandling. Därför är det viktigt som utgångspunkt för vidare diskussioner att redogöra för vad som menas med kvalitet i en korpus och vilka krav som finns när en korpus skapas. Att tolka innehållet i en korpus är inte trivialt. Vid användning av korpusar som utgångsmaterial vid skapande av ordklasstaggningsverktyg krävs helst ett material som är disambiguerat. Därför definieras även textuell homografi i detta sammanhang och vilken betydelse homografin har.

Nedan följer en mer detaljerad beskrivning av kvalitet och innehåll i korpusar, avsnitt 2.1.2.1. I avsnitt 2.1.2.2 diskuteras betydelsen av storlek på korpusar, med hänvisningar till flera studier som gjorts om ämnet.

2.1.2.1 Kvalitet och innehåll

Vid en diskussion om kvalitet i ordklasstaggade korpusar kanske man i detta sammanhang främst talar om kvalitet i bemärkelsen korrekthet i ordklasstaggningen. Men det kan också röra sig om vilken information som finns representerad för varje graford i en korpus, dvs. innehåll i korpusen. Om den informationen är begränsad kanske korpusen eller taggningen av korpusen inte räcker till för sitt syfte, t.ex. som underlag för att träna ett annat verktyg. Detta beror givetvis på tillämpningsområdet.

Kvalitet

I frågan om kvalitet i korpusar har vissa erfarenheter gjorts inom arbetet med SUC. I det arbete som har utförts med SUC 2.0, har en del fel och inkonsekvenser kunnat identifierats; dock har de flesta fel korrigerats från version 1 till version 2 av SUC. Fel och inkonsekvenser redovisas vidare i avsnitt 2.2.2.1. I övrigt kan det diskuteras hur korrekt en ordklasstaggad korpus egentligen kan bli då de flesta korpusar som finns idag primärt har genomgått en automatisk ordklasstaggning och därefter en manuell analys och korrigering. Eftersom det ofta rör sig om relativt stora korpusmaterial, i de flesta fall större än en miljon ord, tar det för lång tid för att endast en person ska kunna utföra manuell korrigering. Därför fördelas ofta arbetet

på flera personer. Ju fler personer som är inblandade, desto svårare är det att få en enhetlig ordklasstaggning. Dessutom beror det också på vilken erfarenhet och kunskap de inblandade har om morfosyntaktisk taggning. Det finns dock möjligheter att mäta kvalitet i taggning t.ex. mellan olika manuella taggare med hjälp av ett mått 'Kappa-måttet'. Se vidare om detta i avsnitt 5.1.1.

En annan aspekt på kvalitet i ordklasstaggade korpusar är den som finns då en korpus används för ett specifikt syfte, t.ex. som utgångsmaterial vid skapandet av en ordklasstaggare. Genom att sammanställa information från en ordklasstaggad korpus kan frekvensuppgifter fås om hur ofta ett ord taggas med en viss ordklass etc. Den typen av statistiska uppgifter är nödvändiga för att en ordklasstaggare automatiskt ska kunna disambiguera och avgöra i vilket sammanhang ett ord får en viss ordklass och när en annan ordklass är aktuell.

I Prütz (1999, 2002) förs ett resonemang om olika kriterier som gäller vid skapande av korpusar. Prütz jämför ordklasstilldelning i SUC och ett projekt som använt UCP och Svensk Ordbok (1986). Då framgår att UCP har en mycket detaljerad analys vad gäller pronomen, men har ingen information om vissa andra syntaktiska kategorier som t.ex. determinerare eller modifierare. Samtidigt innehåller SUC en annan uppsättning syntaktiska kategorier vilket Prütz menar beror på att syntaktiska och funktionella kriterier måste ha legat till grund för utformningen av korpusen.

Detta resonemang stämmer delvis överens med författarens tidigare antydan om att det är viktigt att klargöra syftet med uppmärkning av en korpus vad det gäller användningsområde och utseende. I olika användningsområden behövs olika mycket syntaktisk/grammatisk information medan det i andra områden som t.ex. informationsextrahering kanske är mer intressant att söka efter olika typer av information i texten och inte syntaktiskt/grammatiska uppgifter kopplade till ord. Då behöver inte ordklasstaggningen vara lika specifik. Om syftet med att uppmärka en korpus däremot är att t.ex. träna en taggare som den i avhandlingen är det viktigt att så mycket morfosyntaktisk information som möjligt finns med i korpusen.

Diskussionen om kvalitet ligger nära den om korpusens innehåll, som beskrivs i följande avsnitt.

Innehåll

Det morfosyntaktiska eller uppmärkningsinnehållet i en text har redan i viss mån diskuterats i avsnittet om kvalitet i ordklasstaggade korpusar. Detta avsnitt kommer därför att fokuseras mer på innehållet i en text t.ex. hur innehållet tolkas, i form av vilken homografi som förekommer. Eftersom homografi är ett centralt begrepp i avhandlingen är det viktigt att särskilja olika typer av homografi. I detta sammanhang avses inte den typen av homografi som kan beskrivas som extern, intern, partiell eller interlingval, utan snarare alla typer av homografi. Här kommer homografi att diskuteras utifrån vilken betydelse den har för textens utseende.

Förutom den uppenbara s.k. potentiella homografin vilken uppstår genom att ett graford har mer än en möjlig ordklasstagg i ett lexikon, kommer här att talas om den homografi i en text som föreligger vid två tillfällen, dels den homografi som finns före disambiguering av ordklasstaggad text och dels den som förekommer efter disambiguering.

Den mest självklara homografin av de två som nämndes ovan, är den som förekommer före disambiguering. Där kommer den potentiella homografin också in eftersom den orsakas av att ett graford har definierats som homograft. Om denna definition används vid en första uppmärkning av en text så blir varje graford tilldelat alla möjliga ordklasstaggar som finns i ett lexikon. På så sätt uppstår homografi i en text före disambiguering. Avsikten med disambiguering (automatisk eller manuell) är sedan att välja rätt tagg för grafordet i en viss kontext.

Därmed kommer vi in på disambiguerad text i en korpus. Optimalt finns då endast en ordklasstagg associerad med varje graford. Detta är dock inte alltid fallet. Av olika anledningar kan disambiguering (främst automatisk) inte genomföras för alla ord. Då uppstår en *resterande homografi*. En av anledningarna till detta kan vara att det rör sig om ett sådant fall där kontexten är otillräcklig för att det automatiskt ska kunna avgöras vilken tagg som är rätt i just den kontexten. Dessutom finns det andra typer av homografi, t.ex. inom ett lemma, som kan vara mycket svåra att lösa, även för en mänsklig disambiguerare. I avhandlingsarbetet med EPOS har det framkommit att den största typen av s.k. resterande homografi beror på otillräckliga disambigueringsregler i disambigueringsverktyget. Detta beror i sin tur på att underlagsmaterialet, dvs. träningskorpusen har varit för liten, eller kanske haft för liten variation i typer av text. Denna information om icke disambiguerade ord kan användas i en felanalys i syfte att försöka förbättra det resulterande disambiguerade textmaterialet.

Homografin i en text efter disambiguering kan emellertid också definieras så att alla belagda exempel av ett graford i en text kan resultera i flera möjliga taggar och grafordet kan på det sättet anses vara homograft i texten. Denna typ av homografi kallas i avhandlingen för *belagd homografi*, och kan i sin tur härledas till potentiell homografi vilket blir mer uppenbart då en ordklasstaggad text används för att automatiskt generera ett lexikon. Då blir den potentiella homografin i lexikonet och den belagda homografin för graford efter disambigueringen, densamma. Den belagda homografin i en text är givetvis lägre, dvs. det förekommer lägre antal taggar per graford än i detta fall med lexikongenerering ur texten än i det vanligare fallet där potentiell homografi har sitt ursprung i ett manuellt konstruerat lexikon, som t.ex. SMDB.

2.1.2.2 Storlek

Hur stor en textkorpus är mäts i antal teckensekvenser (token). Dessa inkluderar både textord och interpunktionsord. Storleken är ett begrepp som är lätthanterligt och därför finns det redan ett flertal studier om dess betydelse i olika sammanhang, framför allt då en korpus används som underlagsmaterial för träning av en ordklasstaggare. Det som undersökts i dessa studier är korpusens storlek i relation till korrektheten i en text som taggats av en taggare tränad på texten. Vad som framkommit generellt av dessa studier är den naturliga slutsatsen: att ju större träningskorpus som används, desto bättre presterar taggaren. Vad som vore intressant är också att se vilken betydelse olika typer av texter har för prestationen. Rent intuitivt borde fler texttyper generera större spridning av belägg som används i träning av en taggare och därmed också minska den s.k. resterande homografin i en text. Men å andra sidan kan en större spridning av belägg också orsaka en minskning av antal belägg för mer frekventa ord och kanske därmed också skapa andra fall av homografi som beror på konflikter som vid disambiguering inte kan lösas p.g.a. för få exempel. Men det kanske kan undvikas om antalet textgenrer ökas samtidigt som korpusens storlek ökas.

Eftersom inga studier om betydelse av en träningskorpus storlek i förhållande till en taggarens prestanda har utförts i avhandlingen kan författaren endast hänvisa till en beskrivning av några av de studier som har utförts i ämnet; se vidare kapitel 6.

2.1.3 Disambigueringsmetod

I detta avsnitt sammanfattas flera olika delar som rör ordklasstaggning. Dels görs en beskrivning av definition, historia om ämnet och uppbyggnad

av ett generellt ordklasstaggningssystem i avsnitt 2.1.3.1. Vidare beskrivs vilka tekniker och metoder som finns och används vid ordklasstaggning i avsnitt 2.1.3.2. Då några studier som beskriver mappningsarbete och metoder för detta inte har påträffats kommer det inte att behandlas vidare i forskningsöversikten utan återges i form av en redovisning av erfarenheter och tillvägagångssätt i avhandlingsarbetet i avsnitt 4.1.

2.1.3.1 Inledning om ordklasstaggning

I avsnitt 2.1.1.1 introducerades ordklassindelning och ordklasser som begrepp. Här kommer en beskrivning av syften med ordklasstaggning att redovisas samt lite historik om ämnet. Därefter ges en kort introduktion i hur ett ordklasstaggningssystem (POS-system) kan konstrueras.

Olika syften med ordklasstaggning

Vid uppmärkning av lexikonord med syntaktiska kategorier, dvs ordklasstaggning redovisas allmänt accepterade definitioner av t.ex. Barnbrook (1996) och Leech och Smith (1999).

Ordklasstaggning är en form av syntaktisk analys. För att göra någon form av analys av denna typ, vilket är ett av målen i avhandlingen, behöver man en detaljerad beskrivning av en modell av språket som ska analyseras och dess relation till texten som ska analyseras; jfr Barnbrook (1996). Innan en analys kan göras måste man bestämma sig för vilken typ av analys man vill göra, vilket i sin tur beror på vad analysen ska användas till. Mer specifikt kan man välja syntaktisk/semantisk analys, yt-/djupstrukturanalys etc. Om man ska utvinna semantisk information eller undersöka semantiska komponenter i en sats krävs en semantisk analys i djupstrukturform. Om målet är att göra ordklasstaggning av en text som beskriver generell syntax på ett icke-semantiskt plan behövs en syntaktisk analys. Den ska bl.a. innehålla information om en grammatisk kategori plus viss morfologisk böjningsinformation. Därmed uppstår ett behov av ordklasstaggar. Den grammatiska kategorin behövs för att kunna tilldela ett ord rätt klass och information om böjning behövs i ett senare stadium, t.ex. för att disambiguera homografer. Om taggarna i den syntaktiska analysen hade innehållit information om satsdelar hade ordets funktion i stället beskrivits, vilket kan anses vara ett nödvändigt första steg vid semantisk analys.

Leech och Smith (1999) beskriver ordklasstaggning som ett av de mest populära sätten att lingvistiskt annotera text idag. Populariteten beror dels på att det har forskats mycket inom området och att det är möjligt att på

automatisk väg tagga stora texter med relativt hög korrekthet. Dels beror det på att ordklasstaggning också anses vara ett nödvändigt steg vid syntaktisk uppmärkning av texter som möjliggör olika typer av mer avancerade bearbetningar i analys av naturligt språk.

I stället för att använda termerna annotering eller taggning, beskriver Leech och Smith (1999) olika nivåer i bearbetning av en sats som *parsning*. I parsning finns flera steg av annotering vilka en text kan genomgå. Allt beror på vad texten ska användas till, dvs. vissa steg kan utelämnas beroende på tillämpningsområdet. *Syntaktisk* uppmärkning kan göras i form av ordklasstaggning. I nästa steg i parsningen försöker man skapa ett band mellan form och betydelse. Det är ett krav för t.ex. maskinöversättning, informationsextrahering och igenkänning av tal. *Semantisk* uppmärkning innebär just detta. Vid denna typ av taggning markeras ord enligt deras betydelse i kontexten, antingen med hjälp av ett maskinläsbart lexikon eller ett semantiskt nät. Det finns också en annan tolkning, nämligen att det betyder markering av semantiska roller, men denna forskningsinriktning är inte lika utbredd som den som beskriver annotering av ordbetydelse. Det finns ytterligare ett steg som kan göras i parsningen som har med annotering att göra och det är uppmärkning av *diskurs*. Då markeras länkar som beskriver anaforer och antecedenter.

Mycket energi ägnas idag åt att förbättra ordklasstaggning. En av anledningarna till detta är att den allmänna uppfattningen är att de flesta tillämpningar inom Natural Language Processing *(NLP)* och Language Engineering *(LE)*, som t.ex. syntaktisk analys och parsning som vidare begrepp, också kan förbättras med hjälp av taggade texter, t.ex. vid automatisk inlärning, s.k. *machine learning,* se avsnitt 2.1.3. Andra användningsområden för ordklasstaggning är t.ex. ett område som är ett av de mest elementära där ord och ordklasser jämförs, nämligen i konkordansprogram, där man kan söka efter information om tecken, ord och fraser. Även om det är en enkel representation av ord och dess ordklassrepresentation är detta kanske ett av de vanligaste verktygen för olika språkforskare som inte själva arbetar med system för ordklassanalys utan endast studerar resultaten av sådana analyser. Ordbehandlingsprogram är ett annat exempel på en tillämpning där man har nytta av bl. a. ordklasstaggning. De flesta ordbehandlingsprogram har idag kontroller för både stavning och grammatik. För att åstadkomma sådana kontroller är det en fördel att texten är ordklasstaggad. Om man har en grammatisk representation av en sats och t.ex. upptäcker ett adjektiv som är felstavat, så är det en hjälp för användaren att få möjliga förslag på korrekta stavningar som också är adjektiv och inte nomen eller andra ordklasser.

Efter denna korta sammanfattning av den allmäna nyttan med ordklass-

taggning fortsättes med en redovisning av vilka historiska ansatser som ligger till grund för de ordklasstaggningssystem som finns idag.

Historik om ordklasstaggning

Ett av de första kända fallen då det fanns anledning att använda automatisk ordklasstaggning, var på 1960-talet då Brown Corpus skulle annoteras med lingvistisk information. Brown Corpus bestod av en text på en miljon otaggade textord. En dator kunde då användas för att analysera ett textord och hitta ordets möjliga ordklasser. Problemet var att homografer inte kunde disambigueras. De Rose (1991:9) anger i en studie av just ambiguitet i Brown Corpus att ca 48% av alla "word tokens" (textord) förekommer med mer än en möjlig ordklass. Problemet med homografa ord i Brown Corpus fick lösas med ett datorprogram som skulle disambiguera och tilldela varje ord en ordklasstagg. Greene och Rubin (1971) beskriver ett av de första program som skapades för att automatiskt ordklasstagga stora mängder otaggad text, nämligen programmet TAGGIT som använder en tagguppsättning med 78 taggar. Francis och Kucera (1982) beskriver en annan tagguppsättning för syntaktisk analys av Brown Corpus där alla stora ordklassgrupper finns med plus vissa viktiga särdrag. Funktionsord och interpunktion ingår också. Tagguppsättningen innehåller taggar som kan användas för att markera syntaktiska ordklasskategorier snarare än ordklassernas funktionella betydelse. (Jfr skillnaden mellan ordklasser och satsdelar.) TAGGIT-programmet klarade av att tagga korrekt till 77% och övriga 23% taggades manuellt.

Ända sedan dess har en diskussion pågått om hur ordklasstaggning kan förbättras. Av flera ansatser i taggningsprojekt har bland annat i ett samarbete mellan Lancaster, Oslo och Bergen (1978-1983) LOB Corpus utvecklats och ett nytt taggningssystem CLAWS1 (Constituent Likelihood Automatic Word-tagging System), där taggningssystemet baserades på probabilistiska tekniker. Garside (1987:1) anger att CLAWS1 hade en felprocent på 3-4% vid taggningen när samma tagguppsättning som i TAGGIT användes.

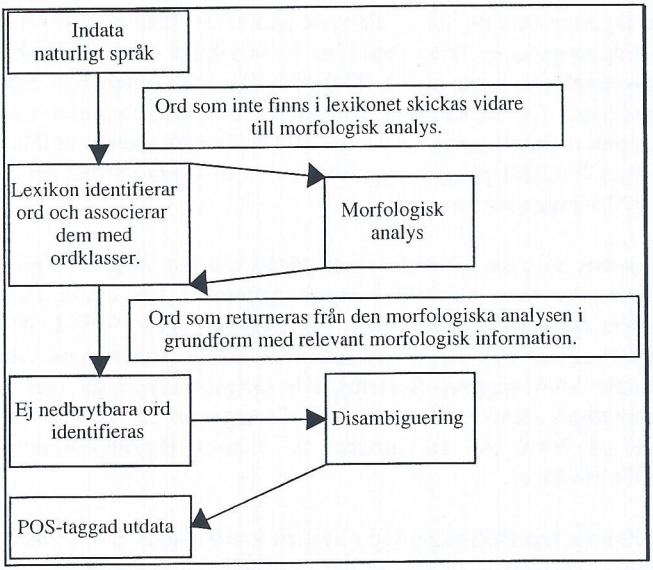
På 1990-talet har POS-taggning blivit en kanske ännu mer livaktig forskningsgren och flera nya metoder har föreslagits, av vilka de viktigaste beskrivs i avsnitt 2.1.3.2.

När det gäller tekniken för ordklasstaggning hävdar Voutilainen (1999) att den största skillnaden är den mellan *lingvistiska* angreppssätt och *dataimplementerade* angreppssätt. Det *lingvistiska* angreppssättet implicerar i princip att en grammatiker skriver taggningsregler för hand. Reglerna baseras ofta på generaliseringar av språket som kan bygga dels på en

subjektiv uppfattning, dels exempel från stora korpusar. Det *data-implementerade* angreppssättet skapar en språklig modell från automatiskt utförda statistiska studier av stora textmängder. Oftast har textmängderna som används för träningen redan taggats av lingvister. Den statistiskt skapade språkmodellen kan representeras på olika sätt, t.ex. som kollokationsmatriser, Markovmodeller, enkla regler eller neurala nätverk. Språkmodellen består oftast av korta ord-/taggsekvenser plus deras frekvenser. Med hjälp av dessa frekvenser väljer taggaren det alternativ som har störst sannolikhet. I avsnitt 2.1.3.2 beskrivs de olika angreppssätten mer ingående.

Uppbyggnad av ett generellt ordklasstaggningssystem

Närmast följer en beskrivning av ett ordklasstaggningssystems principiella struktur, vilket baseras på ett generellt system illustrerat av McEnery och Wilson (1997). De ger en bra bild av de moduler som vanligtvis ingår i ett sådant. Nedan visas en schematisk bild över taggningssystem, figur 3.



Figur 3. Generell beskrivning av taggningssystem, McEnery och Wilson (1997:121).

McEnery och Wilson förklarar skissen över taggningssystem på följande sätt.

Lexikon

Lexikonet kan se ut på olika sätt, men principen är att varje ingång innehåller ett ord plus ett antal möjliga taggar. Ingång betyder lexikoningång, dvs. beskrivning av ett ord i lexikonet och informationen som är kopplad till ordet. Denna typ av lexikon kan antingen skapas för hand eller genereras från stora mängder taggade texter. Lexikonet kan vara av varierande storlek, varvid man kan anta att ju större lexikonet är, desto större är chansen att ett ord får sig tilldelat korrekt tagg. Okända ord som inte finns i lexikonet går vidare till den morfologiska analysen.

Morfologisk analys

Om lexikonet, som beskrevs tidigare, inte innehåller ord med böjningsändelser utan endast grundformer kan en första enkel morfologisk analys identifiera pluraländelser plus tillhörande grundform som finns i lexikonet och därigenom komma fram till den korrekta ordklasstaggen. Detta är svårare att göra för svenska än t.ex. engelska, med tanke på att homografin inom böjningsmorfologin är omfattande. Den morfologiska analysatorn gör oftast en enkel mönstermatchning för att identifiera ändelser. Den kan snarare beskrivas som en ändelseanalys i stället för en morfologisk analys, i och med att den analyserar ändelser och inte egentliga morfem. Däremot används samma tekniker som i morfologisk analys vilket motiverar beteckningen.

Om lexikonidentifieringen inte lyckas identifiera ordet kommer en gissning att äga rum, baserad på tekniker som involverar matchning av ord i den närmaste kontexten och den ändelse ordet har. Här kommer även probabilistiska antaganden in, eftersom det oftast är den vanligaste kombinationen i mönstermatchningsregler som väljs. Ord som inte går att härleda eller analysera vidare till någon grundform och ofta hänger ihop med ett annat ord, med vilket det bildar ett uttryck eller en form, är ej nedbrytbara ord. Dessa kallas mer formellt *syntaktiska idiom* eller *flerordsenheter* (jfr Leechs definition i avsnitt 2.1.1.1), och kan vara en preposition som t.ex. "i samband med" där tre ord tillsammans bildar en preposition, eller "i dag" som tillsammans bildar ett adverb. Dessa grupperas ofta med hjälp av speciella regler för tilldelning av taggar som kan finnas i lexikonet. Flerordsenheter kan härledas från textkorpusar med hjälp av statistiska metoder eller maskinläsbara lexikon, förutsatt att dessa innehåller flerordsenheter.

Disambiguering

Nästa steg är disambigueringen. Efter att en text har passerat lexikon och morfologisk analys har orden fått flera möjliga taggar både ifrån lexikonet och från den morfologiska analysen. Därför behöver man nu disambiguera homografer så att varje ord tilldelas endast en tagg i stället för flera möjliga. Korpusar kan t.ex. användas för att skapa en matris som med hjälp av sannolikhet beskriver hur troligt det är att en tagg följs av en annan. Halliday (1991)[2] menade att människor använder sig av probabilistiskt tänkande när de ska avgöra vilken ordklass som med störst sannolikhet följer efter en annan. Två pionjärer inom detta område är Shannon och Weaver (1949), vilka också beskrivit ords förhållande till varandra och hur de är beroende av varandra på olika sätt. De menade att relativ sannolikhet mellan ord inte är symmetrisk. Härmed ger avhandlingsförfattaren ett litet exempel på detta: Antag att i ett ordpar (X,Y) har X med största sannolikhet taggen A, men Y givet X kan både ha taggarna B och C med lika stor sannolikhet. X har därför ett beroendeförhållande till Y där sannolikheten för X-> A givet Y är större än för någon av Y->B|C givet X.

> Words are produced not as isolated random events, but as part of a coherent structure. This structure is provided by many features, included amongst which is grammar, and it is precisely structure which is required for a stochastic process such as a transition matrix to operate effectively. The existence of structure in language means that words are linked by dependent probabilities. (Shannon och Weaver, 1949 från McEnery och Wilson 1997:123)

Efter framgångsrik disambiguering av homografer har varje ord endast en ordklasstagg. I andra fall, när det inte är möjligt att avgöra vilken tagg som ska användas i ett visst sammanhang, måste fler än en ordklasstagg returneras i utdata.

2.1.3.2 Metoder och tekniker vid disambiguering

Den sista och tredje direkt påverkande faktorn är disambigueringsmetod och -teknik.

I detta avsnitt ges en sammanfattning av metoder och tekniker som används i POS-analyssystem. Beskrivningarna av teknikerna görs dels med avseende på hur materialen som används i teknikerna har framställts och vilka tekniker som används, (lingvistiska eller datorstyrda tekniker), dels

2. Citaten i detta avsnitt är hämtade från McEnery och Wilson (1997:123).

beroende på graden av automatisering vid själva inlärningen när tekniken *machine learning* tillämpas, med avseende på distinktionen mellan *supervised* och *unsupervised* learning (van Guilder, 1995). Vidare beskrivs machine learning samt olika metoder vilka använder sig av inlärningstekniker i detta avsnitt. Inlärning kallas också träning. Att träna en taggare gör man genom att köra ett program med en taggad eller otaggad text som indata. Programmet är styrt så att det extraherar information från texten och lagrar eller bearbetar den på olika sätt beroende på den teknik som används.

Sammanfattningen av tekniker nedan bygger i stor utsträckning på Voutilainen (1999) där han beskriver några olika tekniker som används i en ordklasstaggare, dels i form av automatiska datadrivna tekniker och dels som lingvistiska angreppssätt. Vid beskrivning av tekniker och metoder används den särskilda terminologi som brukas vid utvärdering; terminologin beskrivs i sin helhet i kapitel 5.

Sist i avsnittet redovisas vilka ansatser som har gjorts inom ordklasstaggning för svenska.

Lingvistiska tekniker

Den första storskaliga lingvistiska taggaren kom 1990 (Karlsson 1990). Den byggde på en s.k. regelbaserad grammatik, Constraint Grammar (CG). CG liknar andra system som t.ex. TAGGIT i flera avseenden, även när det gäller disambigueringen, förutom det att CG har en kontextram som sträcker sig ända upp till meningsnivå till skillnad från TAGGIT som endast använder sig av en bigram-modell. När systemet användes för engelska (EngCG), uppvisade systemet i en analys på ca 2000 ord en recall på 99,8 % och en precision på 95,5%, Voutilainen (1992).

Oflazer och Kuruöz (1994), har utvecklat en regelbaserad taggare för turkiska. Den bygger på en morfologisk tvånivåanalys. Den anses ha en korrekthet på 98-99%.

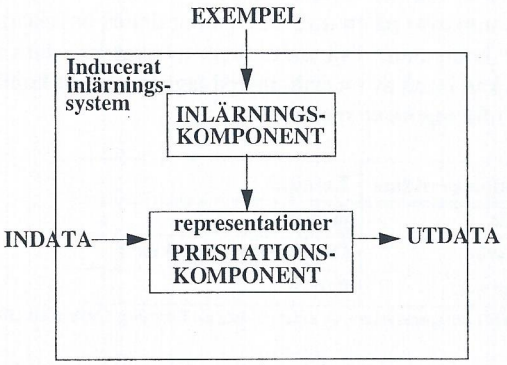
Chanod och Tapanainen (1995a), har utvecklat en finite-state taggare för franska som använder en storskalig morfologisk tvånivåbeskrivning. Grammatiken som används består av tre delar, en funktionsduglig (reliable) del och två heuristiska. De sistnämnda används för lösning av resterande ambiguitet efter resultatet av den första. Reglerna är få och består av sk. transducers. Transducers kan i stort sett jämföras med transformationer. Det handlar alltså om övergångar från ett tillstånd till ett annat. Korrektheten har vid utvärdering uppmätts till ca 97-99%.

Datadrivna tekniker (Machine Learning)

Om uppgiften att skapa probabilistiska regler som behövs i ett analyssystem ses från en människas perspektiv blir det mer komplicerat, och då är det en fördel om man har en stor ordklasstaggad korpus att generera reglerna från. Dels skapas probabilistiska mått som statistiskt sett överensstämmer med innehållet och dels är proceduren relativt enkel. Garside et al. (1987) använde sig av denna teknik när de skapade Constituent Likelihood Automatic Word-Tagging System (CLAWS). Brill och Marcus (1992) är ett senare exempel på några som inte endast använde en korpus för att skapa probabilistiska disambigueringsregler automatiskt utan också för att skapa en tagguppsättning. De använde sig av en "clustering algorithm" för att avgöra hur många olika kombinationer av särdrag det fanns för varje ordklass. De använde också denna genererade tagguppsättning från korpusen för att motivera dess uppbyggnad och omfattning. Förutom material som kan skapas av redan ordklasstaggade korpusar finns det också verktyg som kan skapa material av otaggade korpusar, t.ex. Cutting et al. (1992). I Cuttings taggare sker träningen på otaggade texter med viss övervakning eller genom inlärningsalgoritmer. Denna taggare använder annars samma tekniker som används i CLAWS förutom det att den inte kräver stora taggade korpusar. Detta, säger skaparna själva, gör den språkoberoende och till ett av de enklaste taggningsverktygen som finns idag.

Daelemans (1999) beskriver *machine learning* (ML) som en underdisciplin till *artificiell intelligens* (AI). ML går ut på att använda algoritmer som kan lära antingen genom erfarenhet eller genom att omorganisera kunskap som redan är insamlad. Enligt Daelemans brukar ett inlärningssystem bestå av en komponent som utför arbete efter ett visst mål, en prestationskomponent, och en inlärningskomponent som styr den första komponenten beroende på den insamlade erfarenheten. På detta sätt förbättras prestationen hos inlärningssystemet. Bias är ett begrepp som inom ML används för att beskriva hur ett system är bundet till någon viss domän. Ur lingvistisk synvinkel anger bias hur inlärningen drar nytta av den speciella lingvistiska kunskap som finns i den annoterade träningskorpusen.

Daelemans beskrivning av en generell modell för ett inducerat inlärningssystem (Inductive Learning System) illustreras i figur 4.



Figur 4. Generell modell av ett inducerat inlärningssystem, Daelemans (1999:288).

Nedanstående klassificeringar och metoder som beskrivs vid inlärningssystem återkommer i avsnittet som beskriver hur ML har använts i konstruktionen av EPOS, avsnitt 4.2.2

Vid klassificering av inlärningsmetoder beskriver man vissa egenskaper hos inlärningssystem, som har att göra med strukturen hos informationen och hur den lagras. Exemplen har inriktats på området ordklasstaggning, även om teknikerna kan användas i många andra områden än de med lingvistisk inriktning.

Daelemans anger sex sådana egenskaper nedan:

Övervakning: Hur mycket övervakning som behövs för att ett system ska förbättra sin prestationskomponent och hur detta görs är viktigt att beskriva. De nedanstående metoderna är övervakade tekniker där inlärning sker genom exempel och inte genom fler handskrivna regler.

Representation av indata: Indata kan se ut på olika sätt i ML, men det är ofta par av särdrag och numeriska värden när det gäller det lingvistiska området. Andra områden i vilka man använder sig av ML är t.ex. medicin, kemi, eller mer tekniska ämnen men då består indata naturligtvis inte av lingvistiska särdrag.

Representation av indata	**Exempel**
Indata=(I,N)	I=information (lingvistisk eller annan)
	N=numeriskt värde

Tabell 3. Exempel på representation av indata i Machine Learning, Daelemans (1999:288).

Representation av utdata: Utdata kan i ML bestå av ett tvådelat svar t.ex. ja/nej-beslut, som svar på ett test, t.ex. passar denna ordklasstagg detta ord i just den här kontexten? Två andra typer av svar kan vara en symbolisk kategori som kan bestå av en finit mängd taggar eller en kontinuerlig kategori där svar ofta utgörs av reella tal.

Representation av utdata	Exempel
Tvådelat svar	ja eller nej
Symboliskt svar	[Tagg1,Tagg2,Tagg3 etc.]
Kontinuerligt svar	0.0123

Tabell 4. Exempel på representation av utdata i Machine Learning, Daelemans (1999:289).

Intern representation: Hur informationen som ska läras och representeras beskrivs i inlärnings- och prestationskomponenten, är mycket viktigt att redogöra för. Informationen kan innehålla numeriska värden, t.ex. vid neurala nätverk, eller symbolisk, t.ex. vid semantiska nät, regler eller decision trees. Innehållet i tabellen nedan är exempel och består inte uteslutande av dessa värden utan kan även innehålla andra typer av information.

Typ av information	Typ	Regelutformning	Värde
Symbolisk	Decision tree	(S(NP(PN)),(VP(V)))	0.0123
	Regel	NP -> D, N	0.0456
Numerisk	Neuralt nätverk	Nod1	0.0789

Tabell 5. Alternativ vid intern representation, i Machine Learning, Daelemans (1999:289).

Ökad inlärning (kunskapstillväxt): Inlärningskomponentens kunskap kan expanderas på två sätt. Dels genom att information integreras mellan inlärningskomponenten och prestationskomponenten så att endast ny information lagras. Dels på det sättet som är brukligt i system där informationen inte ökar, sk. *batch learning systems*, som t.ex. i neurala nätverk. Då måste, vid nytillkommen information, all gammal plus den nya informationen läggas in på nytt.

Typ av inlärning	Alternativ
Integrering	1) Ny information som finns sedan tidigare beaktas ej.
	2) Ny information som inte finns vävs ihop med gammal information.
Batch	1) All ny information måste läggas in på nytt liksom den gamla informationen.

Tabell 6. Förenkling av alternativ vid ökad inlärning/kunskapstillväxt, i Machine Learning, Daelemans (1999:289).

Feltolerans: Olika algoritmer är olika känsliga för inlärningsindata som skiljer sig t.ex. genom att delar av informationen saknas. Det är en fördel om algoritmer är toleranta mot sådana fel vid lingvistiskt inlärningsmaterial där kvaliteten kan variera mycket.

Metoder som i olika grad kan klassificeras som *lärande algoritmer* (learning algorithms) (Daelemans):

Table Look-Up är en metod som inte kan klassificeras som "lärande". Vid lagring sparas alla exempel som matas in. Vid användning av lagrade data hämtas matchande obearbetad information med samma utseende som den lagrade informationen vid begäran. Metoden uppvisar dåliga resultat när ett exempel inte matchar indata p.g.a. att det inte finns någon generalisering, dvs. endast exakt match gäller vid sökning i lagrad information.

Användning	**Metod**
Vid lagring	1) Vid lagring läggs all information som kommer som indata in. Bearbetning eller sortering av dessa data förekommer ej.
Vid lookup	1) Vid uppslag eller förfrågan av möjliga POS-taggar för ett ord, anges ett mönster om det finns ett som matchar.
	2) Vid uppslag av okända ord eller ord i okänd kontext ges inget svar om det inte finns något matchande mönster i den lagrade informationen.

Tabell 7. Exempel på användningssätt vid "Table look-up" metoden, i Machine Learning, Daelemans (1999:290).

Case-Based Learning lagrar också alla exempel i en tabell. Men när ny indata ges, hittar metoden ett exempel som liknar indata om exakt match inte finns. Här används bl. a. statistiska och informationsteoretiska tekniker för att definiera likhetsmått, dvs hur lika ett mönster är ett annat.

Användning	**Metod**
Vid lagring	1) Vid lagring läggs all information som kommer som indata in. Indata relateras till annan indata med hjälp av likhetsmått.
Vid lookup	1) Vid uppslag eller förfrågan av möjliga POS-taggar för ett ord, anges ett mönster om det finns ett som matchar.
	2) Vid uppslag av okända ord eller ord i okänd kontext returneras annan "liknande" data som har klassificerats som mer eller mindre lika den data som eftersöks.

Tabell 8. Exempel på användningssätt vid "Case-Based Learning" metoden, i Machine Learning, Daelemans (1999:290).

Rule and Decision Tree Induction använder likheter och skillnader mellan exempel för att konstruera antingen ett "decision tree" eller en regelbaserad representation av exempel från indata. Ett "decision tree" är en datastruktur i vilken noder representerar tester och övergångar mellan noder representerar möjliga svar till tester. Slutnoder (löv) representerar svar på problem. Ett problem löses genom att man följer en väg från rotnoden genom beslutsträdet tills en slutnod har träffats på, dvs ett svar. Konstruktionen används för att avgöra vilken tagg ny indata ska få. Det finns inga specifika exempel, utan det blir mycket generalisering.

Användning	Metod
Vid lagring	1) Vid lagring av information jämförs indata, på liknande sätt som i föregående metod, slutsatser dras om likhet mellan exempel men slutsatser dras också om skillnader, vilka också lagras. Lagring sker antingen i form av decision-trees eller andra typer av regler.
Vid lookup	1) Vid uppslag eller förfrågan av möjliga POS-taggar för ett ord dras hela tiden nya slutsatser utifrån ett träd eller en regel. Inga lagrade regler returneras.
	2) Vid uppslag av okända ord eller ord i okänd kontext kan slutsatser dras om både likheter och skillnader för att ge något svar.

Tabell 9. Exempel på användningssätt vid "Rule and Decision Tree Induction", i Machine Learning, Daelemans (1999:291).

Connectionism, Neural Networks använder exempel för att träna ett nätverk. I sk. "back-propagation"-inlärning, går metoden igenom alla exempel om och om igen för att jämföra den utdata som nätverket föreslår med den korrekta utdatan och vid skillnader ändras vikterna i nätverket för att bättre stämma överens med den korrekta utdatan.

Av de fyra metoderna ovan kan uppslag-i-tabell-metoden, regelinduktion och neurala nätverk räknas till en grupp som använder giriga (greedy) inlärningstekniker. Detta innebär att varje nytt exempel som hanteras vid inlärningen studeras och memoreras på olika sätt. Motsatsen till denna grupp är den som innehåller fallbaserad inlärning, dvs. lata (lazy) inlärningstekniker. Tekniken är "lat" vilket betyder att den endast bearbetar ny indata när den matas med ny information som inte träffats på tidigare. Övrig indata ignoreras.

Sammanfattningsvis kan det konstateras att vid val av giriga eller lata tekniker bör de senare vara en fördel att använda eftersom de är tidsbesparande. Av metoderna ovan är Case-Based Learning lämplig att använda p.g.a. att det är en lat inlärningsteknik och Rule Induction bör med

generaliseringsmetoden vara ett bra komplement. Båda dessa används i EPOS och verkar vara lämpliga som utgångspunkt för hantering av trigram.

Olika datatekniska metoder

N-gram bygger på en modell där ord och taggar representeras som kedjor med sannolikhetsmått. Dessa används för att med en given längd på kontexten avgöra vilket alternativ som är mest troligt, t.ex. vid disambiguering. CLAWS1 är ett exempel på ett system som använder denna teknik. Den sannolikhetsformel, Ekv (1), som används i CLAWS1 för att kunna skapa en balans mellan lexikal och kontextuell sannolikhet, Marshall (1987), är;

$$P = \frac{\textit{Frekvens för taggsekvensen(A,B)}}{\textit{Frekvens för tagg A} * \textit{Frekvens för tagg B}}$$

Sannolikhetsformel från CLAWS1 (Marshall, 1987 i van Halteren, 1999:12). Ekv (1)

Taggsekvensen (A,B) betyder en tagg A följt av en tagg B. Syftet med denna teknik är att automatiskt kunna skapa en uppsättning kontextuella regler av varierande längd (Church, 1988). Reglerna kan beskrivas som lokala och innehåller sannolikhetsmått. En förutsättning för regelgenereringen är tillgång till ett manuellt annoterat inlärningsmaterial.

Metodiken med *datorstyrda lokala regler* använder lokala regler skapade automatiskt från taggade texter. Den bygger på en algoritm utvecklad av Hindle (1989). Denna teknik användes första gången framgångsrikt i systemet TAGGIT. Reglerna har formen;

[PREP + TENSE] = TENSE [N+V]

Exempel 1. Regelbeskrivning från TAGGIT, Hindle (1989).

Detta är ett exempel på en disambigueringsregel i TAGGIT. Regeln innebär att om en sekvens av taggar innehåller en preposition med efterföljande markör i vilken tempusinformation ingår, t.ex. 'to' i engelskan, kommer verbet att väljas i efterföljande kontext. Detta gäller endast om ordet i den efterföljande kontexten kan vara antingen ett nomen eller ett verb.

Reglerna indelas i två grupper där den ena gruppen representerar den vanligaste analysen och den andra representerar ett eller flera undantag. Den andra gruppen är alltid mindre frekvent än den första.

Hidden Markov Models (HMM) är en teknik som i viss mån motsvarar de öppna övergångar som representeras i N-gram, förutom det att övergångarna här är "gömda" (icke-synliga). Övergångarna beskriver t.ex. hur sannolikt det är att en tagg följs av en annan, eller förhållanden mellan en tagg eller en uppsättning taggar, eller en tagg som satsinledare. Alla regler innehåller också information om sannolikhet. Fördelen med HMMs är att det enda som behövs för att träna en taggare är ett lexikon och otaggad text (Cutting et al. 1992). Men vid användning av en taggad text stiger korrektheten (se beskrivning av korrekthet i avsnitt 2.2.2.3) till ca 95-97%. Användning av HMM-tekniken är den vanligaste för taggningsändamål (Manning och Schütze 1999). Taggare finns för de flesta stora europeiska språk.

Neurala nätverk är en teknik där nätverk med flera lager av noder används i taggare. Noderna kan tolkas som en motsvarighet till regler och nätverket kan betraktas som länkningar mellan dessa regler. Noderna får olika vikter vid träning där de har ett indatavärde och ett utdatavärde. Viktningen, dvs. balanseringen av vikterna, fortsätter tills korrekt utdata produceras. Vid träning på en taggad text kopieras sannolikhetsvärdena av förekomster från den taggade texten till nätverket. Schmid (1994) har utvecklat en taggare med denna teknik och korrektheten uppmättes till ca 94%.

Fallbaserade taggare använder en dataimplementerad inlärningsteknik som bygger på en disambiguering med hjälp av lokal kontext och analys av ord. Inlärningskomponenten bygger upp en mängd av fall (en bas). Därefter löses ambiguitet genom att homografer (fall) med närmaste eller mest lika motsvarighet (match) används då regler med identiskt innehåll som nya fall saknas (Daelemans et al. 1996). Taggare som använder denna teknik har utvecklats för bl. a. holländska och engelska och prestationen motsvarar den hos HMM-taggare.

Mer specifika metoder som används vid taggning av olika slag är t.ex.;

Memory Based Learning (MBL), en inlärningsteknik som är en variant av fallbaserade taggare och kan användas för att lagra information om olika fenomen i naturligt språk. Tekniken kan t.ex. användas vid ordklasstaggning, då informationen kan bestå av mönster av ordklasser. Den består av två komponenter, dels en inlärningskomponent, som används för att komma ihåg information, och dels en komponent som utför jämförelser mellan ny information och den som redan finns lagrad och drar därifrån slutsatser (Zavrel och Daelemans, 1999).

Transformation Based Learning (TBL), är en teknik som används för att skapa kontextuella och lexikaliska regler som beskriver övergångar (transformationer) där relationer mellan taggar förekommer. Transformationerna är övergångar från ett tillstånd till ett annat; dessa kan vid ordklasstaggning beskrivas som en tagg eller ett val av tagg i en viss kontext. Dessa transformationer genereras automatiskt från korrekt taggade korpusar, Brill (1992-1994).

Maximum Entropy (ME), är en teknik som använder sig av sannolikhetsmodeller. Modellerna gör det möjligt att kombinera olika delar av kontextuella faktorer för att beräkna sannolikheten för att en speciell lingvistisk klass ska uppträda inom en speciell lingvistisk kontext (Ratnaparkhi 1999). Metoden kan ses som en förfinad variant av andra liknande metoder.

Kombination av flera datatekniska metoder

Förutom de specifika metoder och tekniker som har angetts ovan har även andra ansatser gjorts för att utföra ordklasstaggning med så hög korrekthet som möjligt. Denna metod går ut på att använda flera olika metoder och språkteknologiska resurser för att uppnå bättre resultat, Borin (2000). Detta är ett område som det forskas livligt kring. Borin menar att undersökningar inom ett flerspråkigt översättningsprojekt med parallellkorpusar (ETAP), har visat att man kan utnyttja olika resurser t.ex. taggare om de har systematiska skillnader och vända det till en fördel för att förbättra resultatet av en kombination av resurserna. Han påpekar också att skillnaderna ofta kan vara stora när det gäller tagguppsättning, lexikon och taggningsteknologier, men hänvisar till att om innehållet ses som en modulerad kunskapsresurs är skillnader inget problem. Detta menar han gäller även språkspecifika skillnader mellan taggare. Vidare påpekas att eftersom olika resurser t.ex. taggare presterar olika bra i olika sammanhang är det lämpligt att kombinera resurserna.

Liknande resonemang framförs av Megyesi (2001), vilket redovisas vidare i kapitel 6. Megyesi föreslår dock inte explicit sammanslagning av resurser Hon redovisar en studie där det tydligt framgår vilka för- och nackdelar det finns med att använda vissa specifika taggare.

Taggning av svenska

Nedan följer en redovisning av några av de ordklasstaggare och morfologiska analysverktyg som finns för svenska[3], i kronologisk ordning. En sammanfattning av tillgänglig information om dessa redovisas kortfattat. Utvärderingsmått anges i den mån de existerar, dock av varierande kvalitet. En utvärdering är alltid i viss mån subjektiv beroende på vem som har utfört den, vilken tagguppsättning man har använt vilka utvärderingsmått man har valt, hur måtten har definierats och vilken och hur stor testdata som har använts. Ett annat faktum om tillgänglig information är att "äldre" system, dvs. tidigare än 1987, i de flesta fall inte har beskrivits med metod, teknik och utvärderingsmått, utan snarare med vilken typ av dator som har använts, körningstid och programspråk. Detta gör diskussionen om senare utvecklade system något annorlunda än den om tidigare sådana.

Morph, Ejerhed, Bromley, (1986), är ett system för lexikalisk utvinning. Morfologisk analys av svenska utförs i form av lemmatisering. Systemet skapades för att testa psykolingvistiska teorier. Ingen information om utvärdering har hittats.

LPS (Lexikonorienterad Parser för Svenska), Sågvall (1986) & *UCP (Uppsala Chart Processor)*, Sågvall (1987). LPS var en parser som skulle parsa svenska texter med utgångspunkt från olika maskinläsbara lexikon. Det innehöll bl.a. en lemmatiserare. UCP-formalismen beskriver hela den svenska morfologin och användes i LPS-projektet. UCP är ett procedurellt, unifieringsbaserat chart-omskrivningssystem. Den består av en grammatik och ett lexikon. Ingen information om utvärdering har hittats.

MorP (Morphological Parser), Källgren, (1992), är enligt Källgren en snabb, robust och effektiv morfologisk ordklasstaggare för obegränsade svenska texter. Analysen sker på ytstrukturnivå och använder sig av omskrivningsregler i ett programmeringssystem, BETA (Brodda, 1988). MorP har vid utvärdering uppnått ett mått på 91,4% korrekt taggad text. Utvärderingen verkar inte så förtroendeingivande eftersom man gör jämförelser med situationer då vissa okända ord inte finns med.

SWETWOL, Karlsson (1992). SWETWOL är ett ordklassanalysverktyg som bygger på en tvånivåmorfologimodell för svenska (Koskenniemi 1983). Verktyget beskriver svensk böjningsmorfologi, kan lemmatisera

3. Vid undersökningen av existerande verktyg för svenska studerades endast verktyg utvecklade i akademisk miljö, då det för existerande kommersiella produkter oftast varken finns tillfredställande eller allmänt tillgänglig dokumentation.

och innehåller ett lexikon på ca 48000 enheter. Måtten på recall som har uppmätts har varit utmärkta, men för att förbättra precision har man funderat på att använda en form av lokal disambiguering. SWETWOL har vid en utvärdering uppnått mått på 99,7% korrekt taggad text, vilket verkar vara anmärkningsvärt högt.

En stokastisk ordklasstagger för svenska, Cutting (1993). Cutting gjorde ett försök att utvärdera "The Xerox Part-of-Speech Tagger" (XPOST) på svensk text. XPOST ska vara en "praktisk" taggare, i den meningen att den ska vara enkel att använda ("port") med ett nytt språk. Taggaren konstruerades med målet att den skulle ge korrekt utdata, vara snabb, robust och återanvändbar. XPOST användes med Samuelssons (1993) 13 taggar från ursprungliga 259 i Telemankorpusen. Korrektheten uppgavs till 91%.

Morphological Tagging Based Entirely on Bayesian Inference, Samuelsson (1993), är en taggningsansats med en kombination av heuristiska regler tillsammans med lexikalisk information. Man använder alltså inte enbart den extraherade informationen som finns i en träningskorpus, vilket Samuelsson menar gör taggaren helt språkoberoende. Han använder sig av Bayes formel och trigram för syntaktisk representation. 93,15% korrekt analys på kända ord och 92,59% på okända ord, uppnås.

Ment Model, Blåberg (1994), bygger på en formalism för att representera morfologi och utför lemmatisering. Den använder en grammatisk formalism som är influerad av Generalized Phrase Structure Grammar. Modellen har utvärderats som fungerande system, dock inte med värden på korrekthet.

HMM-tagger, Elworthy (1995), har använt en HMM-taggare som tränades på svenska. Två resultat rapporterades. Det ena gällde taggning av text utan okända ord. Vid en jämförelse med olika stora tagguppsättningar fick han ett korrekthetsmått på 94-97% och ju större tagguppsättning som användes, desto högre korrekthet observerades. Detta resultat är anmärkningsvärt, men är korrekt (enligt personlig förmedling med David Elworthy). Det andra resultatet gällde taggning av text med okända ord. Där blev korrektheten 90-92% på ambiguösa ord och ju mindre tagguppsättning som användes desto större korrekthet konstaterades.

A probabilistic tagger for Swedish using the SUC tagset, Åström (1995), är en taggare som bygger på VOLSUNGA-algoritmen (DeRose 88). Den använder version 4.3.1 av SUC:s tagguppsättning och består av 151 taggar. Taggaren använder lexikaliska listor för lexikalisk uppslagning och Swetwol som morfologisk analysator. Taggaren tränades på 16.000 ord från SUC-korpusen, vilken då endast bestod av 300.000 ord. Vid disambiguer-

ing användes bigram- och trigramregler och prestationen uppmättes till 97,5%.

Rule-Based Tagging in Språkbanken, (1996), Johansson Kokkinakis et al. Detta är ett arbete som utfördes i Språkbanken med Brill tagger som tränades på SUC med en tagguppsättning anpassad till MULTEXT efter rekommendationer för svenska. Arbetet innehåller också en post-editeringsmodul och en möjlighet att provtagga texter, båda med gränssnitt i WWW. Värden på recall och precision uppmättes till 94,2% respektive 96,3%.

A Robust and Modularized Lemmatizer/Tagger for Swedish, (1997), Kokkinakis et. al. är ett arbete som utfördes inom EU-projektet AVENTINUS. Brill tagger användes för disambiguering och SAOL11 användes som lexikalisk resurs. Tagguppsättningen var anpassad efter EAGLES standard för morfosyntaktisk taggning av svenska. Som gränssnitt och kommunikationsverktyg användes GATE. Ingen utvärdering som anger korrekthet har utförts.

LexWare, Dura (1998). LexWare är ett verktyg för textindexering. Både lexikon och statistik används vid språkanalys. Texter annoteras med olika typer av information, bl. a. POS-taggar (samma tagguppsättning som i SUC på CD och Parole). En utvärdering utförd på SUC visar nära 98% täckning vid 100% precision.

μ-TBL, Lager (1999). Ett logiskt programmeringsverktyg för TBL (Transformation Based Learning). Verktyget kan användas för att skapa transformationer för ordklasstaggning. Samma teknik som används i Brill tagger. Vid utvärdering av taggning av nominalfraser har ett värde på över 90% korrekthet uppmätts.

GRANSKA, Carlberger, Kann (1999), är ett projekt vid KTH inom vilket man har utvecklat en stokastisk taggare med en Markovmodell. Taggaren taggar 92% av okända ord korrekt och upp till 97% av alla ord korrekt.

Megyesi (2002); träning av taggare med olika metoder. Dessa anges med namn och uppnådd korrekthetsprocent. Memory Based Learning (MB), 92,28%, Maximum Entropy (ME), 93,49%, Transformation Based Learning (TBL), 92,39%, Trigrams'n Tags (TnT), 95,31%. I samband med träning undersöktes också påverkan på prestation av taggare vid olika stora träningskorpusar. Dessutom undersöktes storlek på tagguppsättning och dess påverkan på prestationen.

Efter denna sammanfattning av kända existerande taggare för svenska kan

konstateras att det fortfarande finns ett behov att täcka i fråga om tillgänglighet. De ovan redovisade är i huvudsak egna utvecklingar, då dessa ofta är mer intressanta att beskriva än den uppsjö av träningar som utförts med existerande taggare, t.ex. Brill tagger. Undantaget är två av författarens egna beskrivningar av Brill tagger vilka kan anses motiverade då de bygger på material från Språkbanken och SAOL. Eftersom de existerande systemen är få och de som fortfarande kan användas ännu färre behövs ett system som kan användas fritt av forskare. Inom Språkbanken finns idag texter och vissa sökverktyg tillgängliga. Men mer avancerad bearbetning såsom morfosyntaktisk taggning och grammatisk parsning finns inte tillgänglig. Därför kommer EPOS att användas för att morfosyntaktiskt tagga alla texter i Språkbanken och i framtiden även göras tillgänglig för andra forskare.

2.2 Indirekta faktorer

I avsnittet om indirekta faktorer ges bakgrund till två faktorer som kan anses påverka prestationen på ett indirekt sätt. De är taggуppsättning och utvärderingsmetod.

Hur en taggуppsättning är utformad har stor betydelse för vilket resultat en utvärdering av ett disambigueringsverktyg kan uppvisa. Det gäller både storleken och sammansättningen/innehållet i en taggуppsättning. Liksom undersökningar som gjorts om storlekar på textkorpusar finns flera studier om just vilken påverkan en taggуppsättnings storlek har på prestationen. Detta diskuteras vidare i kapitel 6. Som underlag för en sådan diskussion är en närmare beskrivning av taggуppsättningar och rekommendationer av sådana en förutsättning; se avsnitt 2.2.1.

Den andra indirekta faktorn, som kan anses ha en självklar påverkan på tolkningen av prestationen vid ordklasstaggning, är på vilket sätt utvärdering av resultat bedrivs. Det rör framförallt vilka metoder som används för själva utvärderingen och vilka delar av ett resultat som är relevant att bedöma, vilket i sin tur beror på syftet med en utvärdering. Utvärderingsmått och metoder redovisas mer i detalj i kapitel 5. I avsnitt 6.2 anges närmare hur utvärderingsmetoder kan påverka prestationen på ett indirekt sätt. Som underlag till detta diskuteras vad som kan menas med korrekt taggning och lite historia om utvärderingsmått och dess benämningar i avsnitt 2.2.2.

2.2.1 Taggуppsättning - standarder och rekommendationer

En taggуppsättning är en uppsättning morfosyntaktiska uppmärknings-

taggar eller etiketter som anger ordklass och morfosyntaktiska egenskaper för ett ord. Nedan ges en allmän beskrivning av kriterier för och innehåll i en tagguppsättning, från EAGLES[4] rekommendationer (1999). Vidare redogörs för SUC:s och SMDB:s tagguppsättning separat i avsnitt 4.1.1 och hur de förhåller sig till de angivna rekommendationerna och kriterierna.

I utgångsläget för avhandlingsarbetet förelåg en situation där man på Språkdata hade skapat en egen tagguppsättning efter egna önskemål och behov för att beskriva ord i ett lexikon samt att märka olika förekomster av textord med diverse morfosyntaktiska särdrag. Det lexikon som används i denna studie är SMDB. Den tagguppsättning, som har använts, är en modifierad version av EAGLES. Anledningen till att man har valt att använda just denna tagguppsättning är att det tidigare med framgång har använts i olika taggningsverktyg vid institutionen för svenska språket och att det lämpar sig väl för att beskriva morfosyntaktiska särdrag för textord i SMDB. Den viktigaste anledningen är att tagguppsättningen bygger på en erkänd standard, vilket gör att det lämpar sig väl för att skapa korpusmaterial som kan användas vid samarbete och utbyte med andra forskningsinstitutioner vilka använder sig av samma standard.

Den tagguppsättning som används i denna avhandling är utformad enligt den av EAGLES gjorda standardiseringen för morfosyntaktisk taggning av ett stort antal europeiska språk. EAGLES startade 1993 och har sedan dess arbetat med att utveckla standarder som bör följas av alla som arbetar med ordklasstaggning[5] och språkteknologi i någon form. Genom att använda ett standardiserat taggningsformat kan material och resultat lätt bli tillgängliga för andra forskare. Genom standardiseringen blir det också möjligt att enkelt integrera verktyg med lexikon och grammatiker och få kompatibilitet i parsning.

I de rekommendationer som EAGLES har utarbetat finns det när det gäller tagguppsättningar dels *obligatoriska, rekommenderade attribut och värden* och dels *speciella extensioner.* Exempel på dessa finns i tabell 10-12. De *obligatoriska* attribut och värden som anges är huvudkategorierna i de största ordklasserna vilka bör finnas med. Exemplen i tabellerna nedan är beskrivna för engelska med undantag av den sista tabellen, där ett exempel har angetts för danska, vilket likaväl kunde gälla för svenska.

4. Expert Advisory Groups on Language Engineering Standards (EAGLES), 1999.

5. I avhandlingen har EAGLES-standarden använts, men det finns även andra standarder som t.ex. TEI och MULTEXT.

Nr.	Ordklass				
1.	N [noun]	2.	V [verb]	3.	AJ [adjective]
4.	PD [pronoun/determiner]	5.	AT [article]	6.	AV [adverb]
7.	AP [adposition]	8.	C [conjunction]	9.	NU [numeral]
10.	I [interjection]	11.	U [unique/unassigned]	12.	R [residual]
13.	PU [punctuation]				

Tabell 10. Obligatoriska attribut och värden, ur EAGLES rekommendationer (1999).

Med *rekommenderade* attribut och värden, Tabell 11, menas värden sådana att om särdragen finns med i det språk som ska representeras så bör dessa attribut finnas med.

Nr.	Attribut	Värden				
(i)	Type:	1. Common	2. Proper			
(ii)	Gender:	1. Masculine	2. Feminine	3. Neuter		
(iii)	Number:	1. Singular	2. Plural			
(iv)	Case:	1. Nominative	2. Genitive	3. Dative	4. Accusative	5. Vocative

Tabell 11. Rekommenderade attribut och värden, ur EAGLES rekommendationer (1999).

Något som bör observeras är att rekommendationerna dock inte gäller genomgående för alla språk. Det finns vissa undantag som kallas *speciella extensioner* (Tabell 12), där språkspecifika attribut och värden anges. Man tillåts också göra vissa ändringar efter behov. För varje attribut finns dessutom ett alternativt värde som kan anges som "0", vilket betyder att inga bestämda värden är tillämpliga.

Nr.	Attribut	Värden			
(vi)	Definiteness:	1. Definite	2. Indefinite	3. Unmarked	[Danish]

Tabell 12. *Speciella extensioner.* Valfria språkspecifika attribut och värden, ur EAGLES rekommendationer (1999).

Förutom de ovan nämnda rekommendationerna bör även utseendet av taggarna i en tagguppsättning diskuteras. Det är viktigt både ur mänsklig och maskinläsbar synvinkel. Leech (1997 i Garside et al.) anger allmänna riktlinjer och kriterier för hur en tagguppsättning ska utformas och användas. Han menar att den tagguppsättning som används vid uppmärkningen av en korpus ofta blir en kompromiss mellan vad som är lingvistiskt önskvärt och vad som är programmeringsmässigt praktiskt möjligt. Eftersom en taggare endast undersöker den närmaste kontexten vid disambiguering av ett ord kommer morfosyntaktiska kategorier i en ordklass inte

att korrekt kunna appliceras med hjälp av en taggare om den inte kan hitta nödvändig information för detta i närkontexten. Det är därför lönlöst att använda dessa särdrag i taggen.

Leech anger tre kriterier för hur namnen på taggarna ska utformas:

1) *Conciseness* (kortfattad/koncis). Korta namn är oftast mer praktiskt att använda än långa omständliga beteckningar. Detta kriterium är vad gäller beskrivning av taggar en motsättning till nästa kriterium, tydlighet.

2) *Perspicuity* (tydlighet). Namn som lätt kan tolkas är mer användarvänliga än namn som inte kan tolkas. Därför är det bäst att använda namn som Preposition i stället för t.ex. "IN", (exempel från Penn Treebank project, Marcus et al. 1993) .

3) *Analysability* (analyserbarhet). Namn vilka kan separeras i logiska delar är att föredra både för den mänskliga användaren och för maskinläsbarheten. Ett exempel är "NP1", vilket kan analyseras som (N=nomen, P=egennamn (proper) och 1=singularis).

De ovan nämnda kriterierna är logiska och självklara krav på hur namn på taggar bör utformas. Det är dock inte alltid lätt att hitta en balansgång mellan vad som är funktionellt och vad som är praktiskt möjligt. I taggning med EPOS har standarden EAGLES följts.

Två andra kriterier som normalt används för att ange hur en tagguppsättning ska konstrueras är, enligt David Elworthy (1995), ett externt och ett internt. Det *externa* kriteriet är att tagguppsättningen måste kunna beskriva de lingvistiska data som ska representeras vid taggning, t.ex. syntaktiska eller morfologiska. Det *interna* kriteriet innebär att taggningen görs så effektiv som möjligt. Elworthy har gjort en studie som beskriver det interna kriteriet i en tagguppsättning. Undersökningen beskriver hur variation i tagguppsättning påverkar prestationen. Resultaten visade att ju mer specifik tagguppsättning som användes, desto större blev precisionen på disambiguering av homografer när en text inte innehöll några okända ord. När texten däremot innehöll okända ord blev resultaten de motsatta. Avsikten var inte att få högsta möjliga precision för ett visst språk, utan att jämföra skillnaden i prestation.

När det gäller förhållandet mellan tagguppsättning och prestation har den tagguppsättning som används vid ordklasstaggning givetvis stor betydelse för prestationen hos en taggare. Följande citat visar att man utifrån en tagguppsättnings storlek inte kan härleda prestation, men allmänt gäller att ju mer specifik information man väljer att representera i en tagguppsätt-

ning, desto sämre prestation får taggaren.

> "A tagger with a correctness of 95 percent with tagset X might well be better than a tagger with a correctness of 97 percent with tagset Y. This is because the tagger performance must be measured against the difficulty of its task.[...] Ambiguity in some features (e.g. tense) can often be resolved with a small context window. For other features (e.g. transitivity) much more information is needed. As a consequence, presence of such features in a tagset usually leads to lower correctness scores. Note that the size of the tagset only has an indirect influence and cannot be used to express the difficulty of the tagger's task." (van Halteren 1999, s. 87).

Vid jämförelse av olika ordklasstaggare tas det ofta inte hänsyn till att storleken på den tagguppsättning som används har viss påverkan på prestationen. Framför allt påverkar de morfosyntaktiska särdrag, som finns representerade, utgången. Särspråklig representation i särdrag har också stor betydelse. Därför bör endast prestation med samma metoder mellan olika språk jämföras, eller åtminstone bör olika taggare ha lika stora tagguppsättningar samt representera likvärda morfologiska särdrag för att rättvisa jämförelser ska kunna genomföras (jfr. van Halterens kommentar ovan).

2.2.2 Utvärderingsmetod

I detta avsnitt ges en bakgrund inför vidare diskussioner om utvärderingsmetod såsom en indirekt påverkande faktor på en ordklasstaggare och dess prestanda. När man pratar om taggning och prestation kan en bra inledning till detta vara vad som egentligen kan anses vara korrekt taggning, se vidare avsnitt 2.2.2.1.

Därefter ges en bakgrund till olika termer och begrepp som används inom utvärdering i språkteknologi. Det är nödvändigt då olika begrepp kan betyda olika saker och används på olika sätt av olika forskare. Utvärderingsmåtten behandlas sedan närmare i avsnitt 2.2.2.3.

2.2.2.1 Vad är korrekt taggning?

Vad eller vem som avgör om en text är korrekt taggad kan bero på en rad olika faktorer. Dels beror det på vilket lexikon man har att utgå ifrån och vilka taggar som finns representerade där för olika ord, dvs. vad en skapare av ett lexikon anser att ett ord kan ha för möjliga taggar. Dels finns en

subjektiv uppfattning om vad som kan anses vara rätt ordklasstagg i olika sammanhang. Dels är felaktigheter och inkonsekvens i taggning en tredje faktor som bör kunna räknas till faktorer som avgör om en text är korrekt taggad eller ej. En text bör därför kunna kallas "korrekt" med avseende på ett visst antal kriterier, dvs. beroende på vilket lexikon som används, vilket val av taggar som kan anses vara normalfallet enligt grammatiska källor (t.ex. SAG) och, enligt samma källor, en viss grad av konsekvens.

Några exempel på de ovan nämnda tre olika fallen då korrekt taggning kan diskuteras ges nedan. Exemplen i texten är hämtade från avhandlingsarbetet och den mappade versionen av SUC 2.0;

1) Ett exempel på skillnader i lexikon, är orden "få" och "många"[6], vilka kan betraktas som pronomen eller adjektiv. I SUC och SWETWOL har man ansett att de är adjektiv medan de finns representerade som pronomen i SAOL/SMDB och NFO. I skapandet av SMDB har man valt att följa den tryckta versionen av SAOL när det gäller val av ordklasstaggar och i skapandet av SAOL har man i sin tur gått på den ordledsuppdelning som finns i NFO4. I SAG (§71, s. 232) beskrivs skillnader mellan adjektiviska pronomen och adjektiv på följande sätt;

> I böjningen visar de flesta pronomen oregelbundenheter i struktur eller funktion i jämförelse med adjektivböjningen. Till skillnad från adjektiven kan adjektiviska pronomen normalt inte kompareras. Emellertid kan inte alla adjektiv kompareras, och det finns ett fåtal pronomen som kan kompareras (kvantitetspronomenen *många, få, mycket, litet...*), SAG (§71, s. 232).

Vidare diskussion om lexikaliska material och dess uppbyggnad följer i avsnitt 3. Lexikaliska och textuella material.

2) Ett annat exempel som gäller subjektiv uppfattning i ordklasstaggningssammanhang är t.ex. proprier. I SUC v. 1.0 taggades inte alla proprier, vissa utelämnades. I SUC v. 2.0 taggades proprier med en speciell namntagg. I Språkbankens taggning med SMDB är målet att tagga proprier med den specifika informationen på varje token men också med markering av större sammanhängande ordsekvenser då de tillsammans bildar ett egennamn. Se nedanstående exempel från SUC 2.0, Tabell 13.

6. Exempelidé från Martin Gellerstam.

Korpusmaterial	Ord	Taggningsformat
SUC 1.0	Gamla testamentet	JJ POS UTR/NEU SIN DEF NOM NN NEU SIN DEF NOM
SUC 2.0	Gamla testamentet	<NAME type=Work> JJ POS UTR/NEU SIN DEF NOM NN NEU SIN DEF NOM </NAME>
SMDB	Gamla testamentet	NPNSND

Tabell 13. Exempel "gamla testamentet" från SUC 1.0, 2.0 och SMDB med olika taggning.

3) Exempel på det sista fallet då man inte kan tala om korrekt taggning är fel och inkonsekvenser. Till fel räknas de exempel på analys som endast förekommer en gång och inkonsekvenser de som förekommer mer än en gång.

Exempel på fel i SUC v.2 är;

... i/PP Machakos/PM NOM ,/MID 70/RG NOM ***kilometer/NN UTR PLU IND NOM*** sydost/JJ POS UTR/NEU SIN/PLU IND/DEF NOM om/PP ...

... upp/AB till/PP 70/RG NOM ***kilometer/NN UTR SIN IND NOM*** och/KN en/DT UTR SIN IND ...

Exempel 2. Ordet "kilometer" med olika numerusinformation i SUC 2.0.

I exempel 2 ska det vara annat numerus i det andra fallet med kilometer.

... omfördelningar/NN UTR PLU IND NOM i/PP finansplanen/NN NEU PLU DEF NOM ...

Exempel 3. Ordet "finansplanen" ska vara NN UTR SIN DEF NOM i SUC 2.0.

... väg/NN UTR SIN IND NOM Stockholm/PM NOM *får/NN NEU SIN IND NOM* nästa/JJ POS UTR/NEU SIN/PLU IND/DEF NOM ...

Exempel 4. Ordet "får" är nomen i stället för verb i SUC 2.0.

Exempel på inkonsekvenser i SUC v. 2.0 är;

... plikt/NN UTR SIN IND NOM att/IE ***ta/VB INF AKT vara/AB*** på/PP ...

... förmåga/NN UTR SIN IND NOM att/IE ***ta/VB INF AKT vara/NN NEU SIN IND NOM*** på/PP ...

Att/IE ***ta/VB INF AKT vara/VB INF AKT*** på/PP barnen ...

Exempel 5. Frasen "ta vara på" med tre olika taggar på 'vara' i SUC 2.0.

tog/VB PRT AKT ***en/DT UTR SIN IND av/PP*** sina/PS UTR/NEU PLU DEF trasor/NN UTR PLU IND NOM

är/VB PRS AKT ***en/PN UTR SIN IND SUB/OBJ av/PP*** bromsklossarna/NN UTR PLU DEF NOM i/PP

av/PP ***en/RG UTR SIN IND NOM av/PP*** intervjupersonerna/NN UTR PLU DEF NOM ,/MID

Exempel 6. Frasen "en av" med tre olika taggar på 'en' i SUC 2.0.

Andra exempel på problem och svårigheter som t.ex. rör översättning av SUC:s taggуppsättning till SMDB:s taggуppsättning finns i avsnitt 4.1.3.

De ovan nämnda faktorerna har påverkat taggningsarbetet i avhandlingen på olika sätt. Dels har det påverkat mappningsarbetet av SUC v. 2 med SUC-taggar och SMDB-taggar och dels har det påverkat den datamaskinella träningen, eftersom den görs på SUC-korpusen, och de disambigueringsregler vilka är utdata från träningen. Därmed påverkas hela disambigueringen av homografer där dessa regler används. Till sist uppstår även problem i utvärderingen, då en mappad bit av den ursprungliga SUC-korpusen används som referenstext och samma ouppmärkta del taggas på nytt med hjälp av SMDB. Då märks åter de skillnader i lexikon som finns i SWETWOL/SUC och SMDB/SAOL.

2.2.2.2 Bakgrund om termer och begrepp inom utvärdering

Som utgångspunkt för avsnittet om utvärdering av taggning av texter med EPOS i kapitel 5, kommer här att redogöras för elementära begrepp som används inom språkteknologi vid utvärdering.

Lite historisk bakgrund ges först om dessa begrepp och dess ursprung. Därefter följer definitioner av olika mått som frekvent används inom utvärdering, i avsnitt 2.2.2.3.

Termerna som används i utvärderingsmåtten vilka anges i avsnitt 2.2.2.3 nedan som t.ex. 'korrekthet' har sitt ursprung i Information Retrieval (IR). Inom IR behövdes mått för att ange hur mycket information av det som eftersöktes som faktiskt hittades och hur mycket som var felaktiga identifieringar. För beskrivning av de utvärderingsmått som redogörs för i avhandlingen se t.ex. van Rijsbergen (1979), Frakes och Baeza-Yates (1992), Grossman och Frieder (1998) och Manning och Schütze (1999).

Utvärderingsmåtten definieras olika beroende på vad det är som ska mätas. Exempel på flera av de utvärderingsmått som används i avhandlingen, beskrivna för att användas inom ett annat språkteknologiskt område, ges

under avsnittet om 'korrekthet' i avsnitt 2.2.2.3.

2.2.2.3 Definition av utvärderingsmått

Inför utvärderingen i kapitel 5 ges nedan en översikt av frekvent använda utvärderingsmått. Det finns även andra sätt att mäta kvaliteten som prestation. Vid utvärdering av SUC användes t.ex. en indelning av feltyper i 'Categorial errors' och 'Feature errors' (Källgren 1992). I vissa fall bedöms kända och okända ord separat (Elworthy 1995). Utvärdering kan vinklas så att den endast speglar vissa delar som presterats. Ofta anges inte vilka texter som använts. Mycket knapphändig information ges vanligtvis om utvärderingen. Enligt författarens mening bör all tillgänglig information redovisas för att en rättvis och korrekt bedömning ska kunna utföras. Detta gäller förutom den traditionella informationen om *precision* och *recall* även information om taggуppsättningsstorlek, innehåll i taggуppsättning, textstorlek, texttyp, homografifrekvens i lexikon och homografifrekvens i texten. Det är också viktigt att välja rätt data att analysera för att på bästa sätt beskriva prestationen av en analys.

Nedan kommer de traditionella utvärderingsmåtten, som t.ex. precision och recall att definieras eftersom de kommer att användas för att bedöma kvaliteten av ordklasstaggningen.

Översikt över utvärderingsmått

Att utvärdera en analys av något slag är nödvändigt för att få ett bra begrepp om kvaliteten. Beroende på vad som ska bedömas kan olika typer av metoder för utvärdering användas. För det första används fyra mått för att bedöma prestationen av författarens analys, nämligen recall, precision, F-measure och 'antal taggar per token'. För det andra görs en indelning av felen i två grupper: kategorifel och morfologiska fel. För det tredje föreslås en ny metod för att beräkna hur reguljär prestation förhåller sig till homografifrekvens i lexikon och text. Detta är speciellt viktigt då SMDB bör kunna anses vara det svenska lexikon som innehåller mest homografi idag.

Måtten recall och precision är två av de mest frekvent använda inom utvärdering av verktyg för analys av naturligt språk och korpusbearbetning och kan beskrivas på följande sätt med avseende på hur man t.ex. mäter prestation av morfosyntaktisk taggning av teckensekvenser. Observera att måtten skiljer sig åt beroende på vilken typ av analys som ska utvärderas. Nedan anges dessa mått enligt en definition[7] från van Halteren (1999) som avser just morfosyntaktisk taggning. Varje definition kompletteras med exempel på taggning av nomen.

Precision

För det statistiska måttet *precision* och dess betydelse inom området ordklasstaggning gäller:

> *Precision* describes how many of the retrieved objects are correct matches for the user's query. For tagging, this means that precision measures how many of the tokens tagged X are tagged X correctly. ...precision is the number of correct token-tag pairs that is produced, divided by the total number of token-tag pairs that is produced, van Halteren (1999:82).

Formel för uträkning av precision, Ekv (2), blir:

$$Precision = \frac{Antal\ korrekta\ enheter}{Antal\ identifierade\ enheter} * 100$$

Formel för uträkning av precision, van Halteren (1999:82). Ekv (2)

Ett exempel: antag att analysen gav 100 identifierade nomen (A). Av dessa är 80 nomen korrekta (B). Precisionen blir då (B/A)*100, dvs. (80/100)*100=80%.

Recall

För det statistiska måttet *recall* och dess innebörd inom ordklasstaggning gäller:

> *Recall* describes how many of the objects matching the user's query are retrieved. For tagging this means that recall measures how many of the tokens that ought to be tagged X are indeed tagged X. ...recall is the number of correct token-tag pairs that is produced, divided by the number of correct token-tag pairs that is possible, van Halteren (1999:82).

Formeln för uträkning av recall, Ekv (3) blir:

$$Recall = \frac{Antal\ korrekt\ taggade\ enheter}{Totala\ antalet\ enheter} * 100$$

Formel för uträkning av recall, van Halteren (1999:82). Ekv (3)

7. Det måste påpekas att det finns flera definitioner av precision och recall i olika typer av litteratur inom korpuslingvistik. Att det finns mer än en kan bero på att man i vissa fall inte har observerat att olika mått bör användas i samband med utvärdering av olika analyser. Författaren har valt att följa van Halterens definition eftersom den verkar mest trovärdig. Det är dessutom den enda mått som författaren har hittat exakt samma beskrivning av i annan relevant litteratur.

Ett exempel: antag att det totala antalet korrekta nomen är 90 (A) och 80 av dessa nomen har taggats korrekt (B). Det innebär att recall då blir (B/A)*100, dvs. 80/90*100=88,9%.

F-measure

Måttet *F-measure* definieras av van Halteren (1999) som (Recall+Precision)/2. Jämför också beskrivningen av F-measure på längre fram i avsnittet från Message Understanding Conference 1998. Där anges att måttet kan viktas beroende på vilken betoning som ska läggas på antingen precision eller recall.

Formel för uträkning av F-measure, Ekv (4) blir:

$$F\text{-}measure = \frac{Precision + Recall}{2} * 100$$

Formel för uträkning av F-measure, van Halteren (1999:82). Ekv (4)

Ett exempel: Med de tidigare givna måtten på recall och precision blir F-measure ((88,9+80)/2)*100=84,45%.

Tag average/token

> *Tag average/token* is calculated as the total number of tags, divided by the total number of tokens, van Halteren (1999:82).

Formeln för uträkning av 'Tag average per token', Ekv (5), blir:

$$Tag\ average\ per\ token = \frac{Totala\ antalet\ taggar}{Totala\ antalet\ token} * 100$$

Formel för uträkning av Tag average per token, van Halteren (1999:82). Ekv (5)

Exempel: Antag att det totala antalet taggar vid en analys är 140 (A). Totala antalet token är 130 (B). Medelantalet taggar per token blir A/B, dvs. (140/130)*100=1,08 taggar/token.

Korrekthet

> *Correctness* is defined as the number of correctly tagged tokens, divided by the total number of tokens. Other names of the same measure is *success rate*, *percentage of correct tags*, or *score*, van Halteren (1999:82).

Definitionen för korrekthet är samma som den för recall, vilket ger följande formel, Ekv (6):

$$Korrekthet = \frac{Antal\ korrekt\ taggade\ enheter}{Totala\ antalet\ enheter} * 100$$

Formel för uträkning av korrekthet, van Halteren (1999:82). Ekv (6)

Ett exempel: Antag att det totala antalet korrekt taggade enheter är 80 (A). Det totala antalet enheter är 90 (B). Korrektheten blir då (A/B)*100, dvs. (80/90)*100=88,9%

Anledningen till att både korrekthet och recall definieras är att de utseendemässigt ser lika ut, men har skapats med olika syften. Korrekthet skapades för att användas för taggare med endast en tagg per ord, medan recall skapades för att användas för taggare vars utdata kunde innehålla mer än en tagg per ord, enligt van Halteren (1999).

För att illustera skillnaden mellan utvärderingsmåtten beroende på vad som ska bedömas, kan man beskriva precision och recall på följande sätt om bedömningen handlar om *NE* (Named Entity) Identification, (MUC-Message Understanding Conference 1998):

> *Precision*=(Total Correct (Penalty*Partially Correct))/Actual
> Actual=Total Correct+Incorrect+Partially Correct+Spurious
> *Penalty* is an arbitrary number used when there are elements not marked by the automatic process given a manually recognized NE segment.
> *Partially* correct means that two annotations overlap, but are not identical.
> *Spurious* marking means that an annotation by the recognizer does not have any response on the manually annotated text.
> *Recall*=(Total Correct (Penalty*Partially Correct))/Possible
> Possible=Total Correct+Incorrect+Partially Correct+Missing
> F-value[8]=(β^2+1) Precision Recall/β^2 Precision+Recall
>
> β is a parameter encoding the relative importance of recall and precision.

Felgruppering

Andra mått används ibland för att spegla prestation inom en viss kategori

8. F-value är samma mått som tidigare beskrivits som F-measure. Olika definitioner innehåller olika namn på samma mått.

eller liknande. Två indelningar i prestation som gäller olika kategorier exemplifieras nedan.

Källgren (1992) diskuterar *Categorial errors (kategorifel)* och *Feature errors (morfologiska fel)*. I sammanhanget innebär Categorial errors fel som fördelar sig på andra ordklasser än de korrekta. Feature errors gäller fel inom en viss ordklass för ett visst enskilt ord, t.ex. numerusfel. Källgren anger ett mått på korrekthet (95,45%). Därefter uppdelas felprocenten 4,55% på Categorial errors 67,89% och Feature errors 32,1%. I utvärderingskapitlet benämns dessa begrepp som *kategorifel* och *morfologiska fel.* Motsvarande siffror från utvärdering av EPOS 2.0 framgår av tabell 33.

Enligt Källgren finns det vid utvärdering av taggning av SUC betydligt fler kategorifel än morfologiska fel.

En annan mätning som ger en uppfattning om morfologisk analys av ord gäller prestation för *kända* respektive *okända ord.* Elworthy (1995) anger i en undersökning en fördelning av prestation mellan dessa kategorier. En fördel med detta mått är att det är möjligt att få en uppfattning om den morfologiska komponent som hanterar just okända ord, samt att man får en allmän uppfattning om hur bra prestationen är, bortsett från analys av okända ord. I EPOS 2.0 anges dessa mått i tabell 64 och tabell 65.

Vidare beskrivningar av metod, resultat och verktyg som använts vid utvärderingen redovisas i kapitel 5.

3 Lexikaliska och textuella material

I detta avsnitt redogörs för olika lexikaliska och textuella resurser, utifrån ett perspektiv som speglar de funktioner de fyller. Enligt EAGLES rekommendationer för morfosyntaktisk taggning är ett lexikon en nödvändighet. Ett undantag till detta påstående är när ett lexikon i egentlig mening saknas, då genererade lexikon från taggade textkorpusar ofta används. En annan viktig resurs är en textkorpus i de fall då disambigueringsregler skapas automatiskt genom s.k. träning. Motsatsen till detta är manuell regelskrivning. Vanligtvis används morfosyntaktiskt uppmärkta korpusar för sådan träning, men det finns också möjligheter att utföra träning på en ouppmärkt korpus, Cutting (1993).

Nedan följer en redogörelse för de lexikon som används i avhandlingsarbetet i avsnitt 3.1, och av de korpusar som använts, i avsnitt 3.2.

3.1 Lexikon

Lexikaliska resurser kan utformas på olika sätt beroende på hur de ska användas. Ett krav på lexikonen är att de är maskinläsbara. Två exempel på maskinläsbara lexikon är SMDB och delvis NFO[9], vilka båda används för olika ändamål i avhandlingen. SMDB används som lexikalisk resurs dels i mappningen av Stockholm-Umeå Corpus (SUC), dels i en annoteringsfunktion (Cederholm), och dels i korrigering i punktinsatser vid disambiguering. Nusvensk Frekvensordbok (NFO) har dels använts i ett försök att utnyttja statistisk information om de 100 vanligaste homografema och dels som referens vid taggning av pressmaterialet Press65.

3.1.1 SAOL-SMDB

Svenska akademiens ordlista har i avhandlingsarbetet använts i två versioner, version 11 och version 12. Den senare versionen, SAOL12, är den som används för EPOS 2.0, den tidigare användes för EPOS 1.0.

Båda versionerna beskrivs eftersom författaren anser att det har betydelse för utvärderingen då prestationen för de båda versionerna kommer att jämföras. En kort sammanfattning ges dessutom om hur SAOL12 har använts

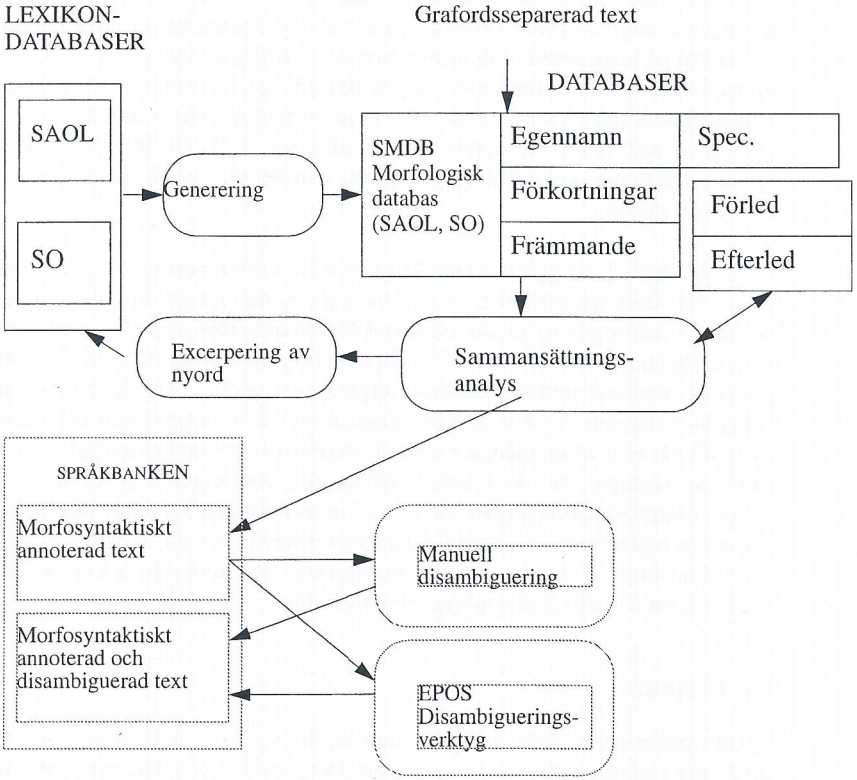
9. NFO fanns tidigare i ett maskinläsbart format men p.g.a. maskinuppdateringar och lägre prioritering på materialet kan det idag inte användas annat än i bokform, med undantag av del 4 vilken fortfarande är tillgänglig.

för att bygga upp SMDB.

SAOL11 har använts för att skapa en lexikalisk fullformsdatabas i Mysql, se Johansson Kokkinakis (1996) samt Kokkinakis et al. (1997). En äldre version av Sture Bergs böjningsklasser (1988) användes för automatisk generering av ett ords alla möjliga böjningsformer. Databasen består av ca 750 000 poster varav ca 80 % är en delmängd av SAOL11. Därtill har ca 6000 insamlade egennamn och ca 95 000 egennamn från Språkbanken lagts till. Den syntaktiska informationen har utökats på 117 adverb och adverbialpartikelinformation har lagts till på 101 adverb. Databasen innehåller information om alla böjningsformer. Det finns också kopplingar till vilket lemma de tillhör. Db-lookup heter den del av EPOS 1.0 som använder det lexikaliska materialet och uppmärker en text med möjliga taggar. Den tagguppsättning som använts består av 282 taggar och bygger på EAGLES rekommendationer vid morfosyntaktisk annotering.

SAOL12 är den senaste versionen av ordlistan (redaktör Martin Gellerstam, bitr. redaktör Sture Berg). SAOL12 har använts för att bygga upp SMDB, Berg et al., (2003). Berg har ansvarat för de ca 300 klasser vilka är grupperingar av orden i SAOL. Dessa klasser ligger till grund för den automatiska genereringen av böjningsformer som representeras i SMDB. SMDB består av ca 1,2 miljoner former. Yvonne Cederholm har varit ansvarig för arbetet med dess uppbyggnad. Cederholm har har också haft övergripande ansvar för utvecklingsarbetet av SMDB. Gellerstam har medverkat i utformningen av ordklasstaggar. I SMDB har en tagguppsättning på 133 taggar använts. Den är snarlik tagguppsättningen i EPOS 1.0 och bygger också på EAGLES rekommendationer. Tagguppsättningen har likheter med den som använts i LE-Parole-projektet, vilken också bygger på EAGLES-standarden.

I figur 5 illustreras systemet runt SMDB, där även framgår vilken koppling som finns till EPOS. För vidare beskrivning av dess innehåll, se rapporten om arbetet med SMDB, Berg et al. (2003).



Figur 5. Systemet runt SMDB, från Berg et al. (2003).

3.1.2 NFO

Nusvensk frekvensordbok (NFO), Allén et al. (1970) är en frekvensstudie som bygger på korpusmaterialet Press 65. Det har resulterat i ett statistiskt material som 1970 var det första i sådan omfattning och utförlighet för svenska. Tidigare arbeten är bl. a. Carita Hassler-Göransson som har sammanställt resultatet av flera olika undersökningar med sammanlagt 500 000 löpande ord (Ordfrekvenser i nusvenskt skriftspråk, 1966). I en studie om tal- och skriftspråksfrekvenser insamlade Allwood (1999) drygt 1 miljon ord. Frekvensbaserade studier utfördes efter automatiskt taggning med en korrekthet på ca 97%[10]. Det finns inte någon senare motsvarighet i NFO:s storlek som gäller skriftspråk. NFO-studien redovisades i form av

en ordbok, som består av fyra delar. Den skapades i syfte att i frekvens redovisa information om materialet dels på grafordsnivå och dels på homografkomponentnivå i första delen. Del två fokuserar på information om orden på lemmanivå. I del tre redovisas ordförbindelser på tre beskrivningsnivåer: ordkombinationer, konstruktioner på huvudnivå och idiom. Konstruktionerna på huvudnivå vilka är grammatiskt intressanta är en delmängd av ordkombinationerna. Del 4 redovisar slutligen ordleden (morfemen) och deras ordbildningsegenskaper och betydelser på fyra olika beskrivningsnivåer.

I avhandlingsarbetet gjordes i punktinsatser i disambigueringsdelen ett försök att använda statistik över de 100 vanligaste homograferna samt den inbördes fördelningen av dessa. Försöket förbättrade inte disambigueringen. Anledningen är sannolikt att de vanligaste homograferna också är de som genererar de mest omfattande disambigueringsreglerna och utgör de mest belagda exemplen i texten. Sådan information om homografer är till störst hjälp då det inte finns många eller helt saknas disambigueringsregler med liknande exempel. Det är i sådana situationer som mycket generella och osäkra disambigueringsregler används. I utvärderingen uppvisar just dessa generella regler sämst resultat i disambigueringen. Om en liknande ansats istället inriktas på mindre vanliga homografer bör man därför kunna förvänta sig ett förbättrat disambigueringsresultat.

3.2 Korpusar

Textkorpusmaterial från två källor har använts i detta avhandlingsarbete med olika syften, dels SUC och dels Press 65. SUC har, efter att ha mappats till en motsvarande korpus med SMDB:s taggar, använts för att med tekniken *machine learning* bygga upp disambigueringsregler i EPOS 2.0. Press 65 har använts i undersökning av resultat av förbättringar av disambiguerad text.

Press 65 och SUC redogörs kortfattat för nedan, i avsnitt 3.2.1 respektive 3.2.2.

3.2.1 Press 65

Press 65 är ett av de textuella material som finns i Språkbanken idag. Korpusen bygger på morgonpress från 1965. Det består av 1 000 669

10. Observera att det vid utvärderingen i fråga hade använts en tagguppsättning med endast 11 taggar.

löpande ord fördelade på 1 387 artiklar av 569 olika författare. Materialet är utdrag från tidningarna Svenska Dagbladet, Stockholms-Tidningen, Dagens Nyheter, Göteborgs Handels- och SjöfartsTidning och Sydsvenska Dagbladet Snällposten, vilket gjordes som ett slumpmässigt urval av olika morgontidningar för att få både en regional och politisk spridning. Artiklarna har också hämtats från olika genrer som t. ex. naturvetenskap, kulturvetenskap, politik och samhällsfrågor, näringsliv, människa och miljö, konstnärlig verksamhet. Tidigare har korpusen använts framför allt som underlag för framställandet av Nusvensk Frekvensordbok på institutionen.

Korpusen är annoterad enligt standarden Corpus Encoding Standard för XML (XCES) vilken bygger på rekommendationer från EAGLES. Det är också det som används i Språkbankens arkivformat, se vidare om detta i Berg et al. (2003). Tidigare har texter uppmärkts med Standard Generalized Markup Language (SGML). I ett första steg har både heterografer och homografer uppmärkts med en annoteringsfunktion med koppling till SMDB (Cederholm). I ett andra steg har homograferna disambiguerats med EPOS. Den tagguppsättning som använts består av 133 taggar. Informationen som representeras i de både uppmärkningarna är tokennummer (tnr), ordklass/ordklasser (msd), lemma, annoteringstyp (atype), disambigueringstyp (dtype), ortografisk information (orth) och normaliserad version (norm) av ett ord. De olika annoteringstyperna är anvisningar om olika källor som informationen hämtas ifrån, t.ex. namndatabas (6,16), årtal (9) eller okänt ord (0). Disambigueringstyperna är information om vilken analys som använts vid disambiguering, t.ex. bigramanalys (12), analys av okända ord (9) eller generell trigramanalys (3). Se exempel 7 med annoterad, ej disambiguerad, text. Se också exempel 8 med disambiguerad text.

```
<ana tnr='0-2' atype='16' msd='NPUSND' lemma='-.-' orth='James Broom Lynne'/>
<ana tnr='3' atype='1' msd='AF0*PN*' lemma='debutera.1' orth='debuterade'/>
<ana tnr='3' atype='1' msd='AF0USND' lemma='debutera.1' orth='debuterade'/>
<ana tnr='3' atype='1' msd='AF0NSND' lemma='debutera.1' orth='debuterade'/>
<ana tnr='3' atype='1' msd='V0ISA' lemma='debutera.1' orth='debuterade'/>
<ana tnr='4' atype='9' msd='MC*PN*' lemma='-.-' orth='1963'/>
<ana tnr='5' atype='1' msd='NCNSNI' lemma='i.1' orth='i'/>
<ana tnr='5' atype='1' msd='R0' lemma='i.2' orth='i'/>
<ana tnr='5' atype='1' msd='S' lemma='i.2' orth='i'/>
<ana tnr='5' atype='1' msd='NCNPNI' lemma='i.1' orth='i'/>
<ana tnr='6' atype='6' msd='NPNSND' lemma='-.-' orth='London'/>
<ana tnr='7' atype='1' msd='NCUSNI' lemma='med.2' orth='med'/>
<ana tnr='7' atype='1' msd='R0' lemma='med.1' orth='med'/>
<ana tnr='7' atype='1' msd='S' lemma='med.1' orth='med'/>
<ana tnr='8' atype='0' orth='The'/>
<ana tnr='9' atype='0' orth='Trigon'/>
```

Exempel 7. Exempel på annoterad text i Press 65.

löpande ord fördelade på 1 387 artiklar av 569 olika författare. Materialet är utdrag från tidningarna Svenska Dagbladet, Stockholms-Tidningen, Dagens Nyheter, Göteborgs Handels- och SjöfartsTidning och Sydsvenska Dagbladet Snällposten, vilket gjordes som ett slumpmässigt urval av olika morgontidningar för att få både en regional och politisk spridning. Artiklarna har också hämtats från olika genrer som t. ex. naturvetenskap, kulturvetenskap, politik och samhällsfrågor, näringsliv, människa och miljö, konstnärlig verksamhet. Tidigare har korpusen använts framför allt som underlag för framställandet av Nusvensk Frekvensordbok på institutionen.

Korpusen är annoterad enligt standarden Corpus Encoding Standard för XML (XCES) vilken bygger på rekommendationer från EAGLES. Det är också det som används i Språkbankens arkivformat, se vidare om detta i Berg et al. (2003). Tidigare har texter uppmärkts med Standard Generalized Markup Language (SGML). I ett första steg har både heterografer och homografer uppmärkts med en annoteringsfunktion med koppling till SMDB (Cederholm). I ett andra steg har homograferna disambiguerats med EPOS. Den tagguppsättning som använts består av 133 taggar. Informationen som representeras i de både uppmärkningarna är tokennummer (tnr), ordklass/ordklasser (msd), lemma, annoteringstyp (atype), disambigueringstyp (dtype), ortografisk information (orth) och normaliserad version (norm) av ett ord. De olika annoteringstyperna är anvisningar om olika källor som informationen hämtas ifrån, t.ex. namndatabas (6,16), årtal (9) eller okänt ord (0). Disambigueringstyperna är information om vilken analys som använts vid disambiguering, t.ex. bigramanalys (12), analys av okända ord (9) eller generell trigramanalys (3). Se exempel 7 med annoterad, ej disambiguerad, text. Se också exempel 8 med disambiguerad text.

```
<ana tnr='0-2' atype='16' msd='NPUSND' lemma='-.-' orth='James Broom Lynne'/>
<ana tnr='3' atype='1' msd='AF0*PN*' lemma='debutera.1' orth='debuterade'/>
<ana tnr='3' atype='1' msd='AF0USND' lemma='debutera.1' orth='debuterade'/>
<ana tnr='3' atype='1' msd='AF0NSND' lemma='debutera.1' orth='debuterade'/>
<ana tnr='3' atype='1' msd='V0ISA' lemma='debutera.1' orth='debuterade'/>
<ana tnr='4' atype='9' msd='MC*PN*' lemma='-.-' orth='1963'/>
<ana tnr='5' atype='1' msd='NCNSNI' lemma='i.1' orth='i'/>
<ana tnr='5' atype='1' msd='R0' lemma='i.2' orth='i'/>
<ana tnr='5' atype='1' msd='S' lemma='i.2' orth='i'/>
<ana tnr='5' atype='1' msd='NCNPNI' lemma='i.1' orth='i'/>
<ana tnr='6' atype='6' msd='NPNSND' lemma='-.-' orth='London'/>
<ana tnr='7' atype='1' msd='NCUSNI' lemma='med.2' orth='med'/>
<ana tnr='7' atype='1' msd='R0' lemma='med.1' orth='med'/>
<ana tnr='7' atype='1' msd='S' lemma='med.1' orth='med'/>
<ana tnr='8' atype='0' orth='The'/>
<ana tnr='9' atype='0' orth='Trigon'/>
```

Exempel 7. Exempel på annoterad text i Press 65.

```
<w n=14>stormaktsblocken<ana><ps>NN<m>NEU PLU DEF NOM<b>stormakts-
block</w>
<w n=15>och<ana><ps>KN<b>och</w>
<w n=16>nedrustningssträvanden<ana><ps>NN<m>NEU PLU IND NOM
<b>nedrustningsträvande</w>
<w n=17>i<ana><ps>PP<b>i</w>
<name type=place>
<w n=18>Europa<ana><ps>PM<m>NOM<b>Europa</w>
</name>
<w n=19>har<ana><ps>VB<m>PRS AKT<b>ha</w>
<w n=20>inte<ana><ps>AB<b>inte</w>
<w n=21>mycket<ana><ps>JJ<m>POS NEU SIN IND NOM<b>mycken</w>
<w n=22>motsvarighet<ana><ps>NN<m>UTR SIN IND NOM<b>motsvarighet
</w>
<w n=23>i<ana><ps>PP<b>i</w>
<name type=place>
<w n=24>Mellanöstern<ana><ps>PM<m>NOM<b>Mellanöstern</w>
</name>
<d n=25>.<ana><ps>MAD<b>.</d>
</s>
```

Exempel 9. Exempel på taggning i SUC 2.0.

SUC 2.0-korpusen har legat till grund för hela träningen av taggaren EPOS. Ur den har hämtats exempel på morfologisk ändelse- och kontextinformation som kan användas vid morfologisk analys av okända ord. Trigram har också extraherats från SUC. Dessa används för att skapa disambigueringsregler vilka används vid homografseparering.

Extra information om homografer har tagits fram genom en samkörning av möjliga homografer i den lexikaliska databasen och både specifika och generella exempel av dessa i SUC-korpusen. I EPOS 1.0 extraherades adverbialpartikelinformation ur korpusen och lagrades för att senare användas vid en specialanalys av just adverbialpartiklar.

Något som skiljer taggningen i SUC från den i t.ex. Press 65 är markering av flerordsenheter. I exempel 10 och exempel 11 nedan ges ett exempel på en flerordsenhet 'i går', vilken kan taggas antingen som 'i går' eller 'i', 'går'. På grund av att den representeras som 'i går' i SMDB kommer den även att representeras så i material taggade med SMDB (t.ex. Press 65). I SUC taggas orden däremot som 'i' och 'går'.

```
<ana tnr='13738-13739' atype='1' msd='R0' lemma='i går.1' orth='i går'/>
<ana tnr='13740' atype='1' msd='NC0000' lemma='morse.1' orth='morse'/>
<ana tnr='13741' atype='1' msd='AQPUSNI' lemma='satt.1' orth='satt'/>
<ana tnr='13741' atype='1' msd='RP' lemma='satt.1' orth='satt'/>
```

```
<ana tnr='13741' atype='1' msd='AQPNSNI' lemma='satt.1' orth='satt'/>
<ana tnr='13741' atype='1' msd='V0ISA' lemma='sitta.1' orth='satt'/>
<ana tnr='13741' atype='1' msd='V0U0A' lemma='sätta.1' orth='satt'/>
<ana tnr='13741' atype='1' msd='AF0USNI' lemma='sätta.1' orth='satt'/>
<ana tnr='13741' atype='1' msd='AF0NSNI' lemma='sätta.1' orth='satt'/>
```

Exempel 10. Exempel på en flerordsenhet taggad med SMDB.

```
<w n=1384>i<ana><ps>PP<b>i</w>
<w n=1385>går<ana><ps>AB<b>går</w>
<w n=1386>morse<ana><ps>NN<m>- - - -<b>morse</w>
<w n=1387>satt<ana><ps>VB<m>PRT AKT<b>sitta</w>
```

Exempel 11. Exempel på en flerordsenhet taggad i SUC.

Vid taggning som i exempel 10 ovan minskar risken för fel vid disambiguering i och med att det homografa ordet 'går' inte behöver disambigueras. Därför är den varianten att föredra. En jämförelse mellan material som skiljer sig åt genom olika representation av flerordsenheter försvåras då antalet taggar per token blir färre i ena materialet. En sådan situation uppstod i utvärderingen av EPOS, när 10% av SUC användes som facit. Den korrekta versionen av SUC mappades till samma tagguppsättning som används i SMDB. Flerordsenheter fanns inte med i den versionen. Den version som taggades av EPOS, vilken var den som skulle utvärderas, innehöll däremot flerordsenheter. Det var inte möjligt att jämföra de två texterna utan något slags facit över flerordsenheter.

4 EPOS: utvecklings- och metodavsnitt

I skapandet av ordklasstaggaren EPOS har vissa beslut angående metod, teknik och tillvägagångssätt tagits. Avsnittet beskriver flera skilda metoder som använts för olika ändamål i avhandlingsarbetet. Det är dels den metod som har använts i mappningsarbetet i avsnitt 4.1, dels disambigueringsmetoden i avsnitt 4.2 och dels vilken metod som använts vid morfologisk analys av okända ord, i 4.2.2.5.

Redogörelserna i detta avsnitt bör vara av allmänt intresse för planerade projekt med liknande inriktning eftersom det ofta kan vara svårt att förutse problem som kan uppstå som gäller material, metoder/tekniker och mål med taggningen. Problem med ordklasstaggning som beskrivs i arbetet är både av mer allmän karaktär och mer språkspecifik karaktär. Det ger en inblick i ett utförligt lexikon som SMDB och ett mycket stort textmaterial som Språkbankens texter.

4.1 Mappningsmetod

En grundförutsättning i de fall då skapandet av en ordklasstaggare sker med hjälp av tekniken *machine learning* är att man redan har ordklasstaggade korpusar att utgå ifrån. Korpusarna används i sin tur för att extrahera ordklassmönster ifrån. Mönstren kan se olika ut beroende på vilka datatekniska metoder som sedan använder dem (avsnitt 2.1.3.2). Dessa mönster bearbetas och bildar tillsammans med statistiska data syntaktiska regler som kan användas i t.ex. disambigueringsarbete. I strävan att få ett uniformt system med korpus, ordklassmönster, regler och annan information som en taggare kan bestå av, bör allt material vara uppbyggt med samma taggуppsättning[11].

Initialt var den enda ordklasstaggade korpus som fanns att tillgå för träning av en taggare SUC-korpusen, vilken är uppbyggd med en egen taggуppsättning. Vid användning av ordklasstaggade korpusar för att skapa en taggare är det brukligt att man delar upp korpusen i två delar. En del kallas *guldstandard* eller *facit* bestående av ca 10% och används i ett slutskede för utvärdering och en annan del på ca 90% används för träning och benämns fortsättningsvis *träningskorpus*. Att använda ett träningskorpusmaterial med annan taggуppsättning medför svårigheter. Inget annat val fanns varför den enda lösningen var att skapa en modul för över-

11. För definition och beskrivning av taggуppsättningar, se avsnitt 2.2.1 och även avsnitt 4.1.1.

sättning av en tagguppsättning till en annan, en s.k. *mappningsmodul*. Mappningen innebär att man utgår från en *källkorpus*, vilken innehåller den ursprungliga korpusen med den text man vill använda för t.ex. träning av en taggare. Texten och dess taggar måste då översättas till en text med andra taggar, dvs. en annan tagguppsättning. Mappningen utförs sekventiellt ord för ord, från vänster till höger. Den nya korpusen kallas *målkorpus*. Svårigheter består bl. a. av att översättningen mellan en tagg och en annan inte alltid består av ett 1:1-förhållande (avsnitt 4.1.3). Målet med mappningen i denna avhandling är därför att skapa en ny korpus med identiskt textmaterial som källkorpusen, men med den tagguppsättning som används i SMDB.

Mappningsarbetet utfördes i den tidigare versionen av EPOS, v. 1.0, (1999), i full skala då alla ingående delar av mappningen ingick och utfördes i en modul. Situationen har förändrats något i EPOS v. 2.0 eftersom mappningen nu utförs inom en modul, men anropar en subfunktion för att hitta möjliga motsvarigheter i taggar (se även Berg et al. 2003).

Mappningsarbetet beskrivs i följande avsnitt: redovisning av de olika tagguppsättningarna i SUC och SMDB/EPOS 2.0 samt EPOS 1.0 (avsnitt 4.1.1). För översättning av taggar i SUC-korpusen till SMDB:s tagguppsättning skapades modulen "suc2saol" (avsnitt 4.1.2). Problem och svårigheter i mappningsarbetet (avsnitt 4.1.3).

4.1.1 Jämförelser av olika tagguppsättningar

I detta avsnitt följer en jämförande undersökning av de tagguppsättningar som behandlas i avhandlingen. Undersökningen bygger på de rekommendationer och kriterier som redogjordes för i avsnitt 2.2.1.

De två tagguppsättningar som är aktuella i denna studie är dels det som SUC-korpusen innehåller och dels det som finns i SMDB. I EPOS v. 1.0 var tagguppsättningen inte identisk med den i SMDB och EPOS 2.0, utan vissa skillnader finns. För att belysa skillnaden i tagguppsättning och dessutom göra en jämförelse av prestationen mellan EPOS v. 1.0 och 2.0 så kommer även v. 1.0 att beskrivas nedan.

SUC (v. 4.3.10)

SUC:s tagguppsättning (v. 4.3.10) består av 153 taggar och skapades med syftet att beskriva ordklass, morfologisk böjningsinformation och lemma. I utarbetandet av tagguppsättningen har man utgått ifrån en tagguppsättning som användes i SWETWOL (Karlsson 1992) samt följt TEI:s[12]

rekommendationer. Innehållet i tagguppsättningen liknar det i SWETWOL men utseendet skiljer sig. Modifieringar har dessutom gjorts, bl.a. i subkategorisering av adverb samt placering av attribut i taggarna. Antal ordklasser uppgår till 23 med en tvåstavig bokstavskod, vilken är en blandning av svensk och engelsk förkortning. Morfosyntaktiska egenskaper, såsom genus och numerus beskrivs med en kod bestående av tre bokstäver. För ytterligare beskrivning av SUC-korpusen, dess innehåll och uppbyggnad, se avsnitt 3.2.2.

I förhållande till EAGLES rekommendationer för morfosyntaktisk taggning uppfyller SUC kraven när det gäller obligatoriska attribut, dvs. ordklassrepresentation. Av de attribut som kallas rekommenderade uppfylls kraven endast delvis. Det innebär att av den inbördes beskrivning som finns inom en ordklass finns inte alla kategorier med. Generellt sett är täckningen godtagbar, men när det gäller t.ex. verb så saknas både särskiljning mellan normala verb och hjälpverb (Ejerhed, 1992), och angivelse av deponens och s-form. Det finns heller ingen noggrannare indelning av adverb. Användning av språkspecifika attribut finns dock, som t.ex. bestämdhet på nomen. Vid en jämförelse med SMDB:s tagguppsättning förekommer fler generaliseringar i SUC-taggarna. Ofta skiljs inte genusbeskrivning åt utan alla tillåtna värden anges, t.ex. NEU/UTR (utläses neutrum eller utrum), eller numerusangivelse SIN/PLU (singular eller plural). Ett konkret exempel är adjektivet "vita", vilket som adjektiv i SUC skulle taggas med en av två möjliga former. I SMDB skulle det kunna få tre möjliga former. Se exempel i tabell 14 nedan;

SUC
vita/JJ POS UTR/NEU PLU IND/DEF NOM duvor/NN UTR PLU IND NOM
vita/JJ POS UTR/NEU SIN DEF NOM korset/NN NEU SIN DEF NOM
vita/JJ POS UTR/NEU SIN DEF NOM duken/NN UTR SIN DEF NOM
SMDB
vita/AQP*PN* duvor/NCUPNI
vita/AQPNSND korset/NCNSND
vita/AQPUSND duken/NCUSND

Tabell 14. Exempel på skillnad i taggning mellan SUC och SMDB.

Notera skillnaden mellan '0' i tidigare exempel och '*', där '0' betyder att värden inte är tillämpliga, och '*' som i exemplen ovan står för vilket som

12. TEI, Guidelines for Electronic Text Encoding and Interchange, http://etext.lib.virginia.edu/TEI.html.

helst av möjliga värden i den position som asterisken förekommer.

Det är tveksamt om SUC uppfyller det första av de tre kriterier som Leech anger för namnen på taggarna i en tagguppsättning. Taggarna och dess attribut med värden har visserligen var och ett för sig kortfattade namn men när alla värden i en tagg sätts ihop blir taggen lång. Därför underlättar det att läsa SUC-taggad text radseparerad för varje token. Det andra kriteriet uppfylls delvis. De flesta taggarna, dess attribut och dess värden kan lätt tolkas och innebörden blir tydlig, men det finns undantag, t.ex. taggen "JJ" för adjektiv. Det tredje kriteriet uppfyller SUC definitivt. I och med att varje attribut i en tagg särskiljs med ett mellanrum är det inga problem att segmentera en tagg i dess ingående delar.

När det gäller Elworthys definition av externa och interna kriterier för en tagguppsättning, uppfyller SUC det externa kriteriet genom att det är möjligt att beskriva de flesta syntaktiska och morfologiska data som ska representeras vid taggning, med undantag av de enskilda fall som angavs ovan. Man kan också tolka detta kriterium på ett sätt som är relativt till det ändamål taggningen ska användas för. Därför kan man tolka kriteriet så att det uppfylls av tagguppsättningen för de ändamål som SUC var menat, nämligen att i största möjliga mån åstadkomma en täckande morfosyntaktisk taggning. Det interna kriteriet, vilket har att göra med effektivitet med avseende på precision i taggning, är varken bättre eller sämre än många andra tagguppsättningar. Det enda man kan konstatera är, att på grund av att vissa attribut inte får specifika värden utan mer generella sådana (se exempel ovan med genus och numerus) blir precisionen högre än för andra tagguppsättningar vilka anger mer specifika värden på attributen. Chansen att tagga fel ökar i stort sett propertionellt till antal möjliga alternativ för värdena på attributen.

SMDB och EPOS 2.0

SMDB:s tagguppsättning (Berg och Cederholm, 2001), består av 133 taggar och skapades för att beskriva de textord som finns i den maskinläsbara, uppblåsta[13] versionen av SAOL 12, vilken innehåller 120000 uppslagsord. Ett mål vid konstruktionen av tagguppsättningen var att; "Den morfosyntaktiska beskrivningen i SMDB ska så långt som möjligt överensstämma med böjningsinformationen i SAOL12." (Berg et al. 2003) Ett ytterligare mål var att beskrivningen skulle ligga nära Nusvensk frekvensordbok (Allén 1970, 1971) när det gäller intern homografi inom ett lemma.

13. 'Uppblåst' betyder att alla möjliga böjningsformer av ett lemma läggs till i t.ex. ett lexikon.

Taggуppsättningen har sitt ursprung i den internationellt erkända standarden EAGLES, dels på grund av tidigare positiva erfarenheter vid institutionen med ordklasstaggningsarbete med samma taggуppsättning och dels för att i framtiden ha möjligheten att utbyta information och ordklasstaggade texter texter med andra användare av samma standard. SMDB innehåller ca 1,2 miljoner former varav 99,98% procent är lexikonord och 0,02% är flerordsenheter. Antal homografa graford är 213193 (24%) och antal heterografa graford är 684316 (76%).

Vid ytterligare en jämförelse mellan SMDB och SUC kan det konstateras att det generellt sett anges mer specifika värden i de attribut som finns i SMDB:s taggуppsättning. I SMDB används ett annat värde, "0", dvs. specifikt värde är ej tillämpligt, i vissa fall.

Eftersom SMDB:s taggуppsättning bygger på EAGLES rekommendationer behandlas inte här hur väl SMDB följer dessa. Av Leechs tre kriterier uppfylls det första. Taggarnas namn för ordklasser innehåller endast en bokstav. Attributen representeras med en bokstav. Vad det gäller det andra kriteriet med krav på tydlighet, ligger SMDB på ungefär samma nivå som SUC. Där är vissa taggar mycket enkla att tolka, samtidigt som vissa andra är svårare, t.ex. "R" för adverb. Här har man gjort ett medvetet val, att sätta kortfattade beskrivningar före tydlighet. Det tredje kriteriet, möjligheten att enkelt separera de olika delarna i taggen, måste sägas vara uppfyllt eftersom varje tagg och dess ingående attribut endast består av ett tecken var.

SMDB:s förhållande till Elworthys två kriterier, kan karakteriseras på följande sätt: det externa kriteriet, där taggуppsättningen måste kunna beskriva de lingvistiska data som ska representeras vid taggningen, uppfylls både vad det gäller syntaktisk och morfologisk information i något större utsträckning än för SUC. Det interna kriteriet, att göra taggningen så effektiv som möjligt med avseende på precision, uppfylls tillfredsställande med tanke på lexikonets storlek och den morfologiska information som finns i varje ordklass.

SAOL och EPOS 1.0

Vid utvecklandet av EPOS v.1.0 användes en taggуppsättning bestående av 282 taggar vilket också var en modifierad delmängd av det ursprungliga EAGLES-taggуppsättningen med viss influens från MULTEXT[14]. Syftet med Multext är att medverka till en standardisering av data, språk-

14. Multilingual Text Tools and Corpora. http://www.lpl.univ-aix.fr/projects/multext/.

teknologiska verktyg och lingvistiska resurser för att maximalt utnyttja återanvändningen av korpusbaserad språkteknologiforskning och tillämpningar. Multext har i viss utsträckning sammarbetat med andra standardiseringsansatser såsom EAGLES och TEI.

Eftersom syftet med v. 1.0 av EPOS var att efter steget med morfosyntaktisk uppmärkning även genomföra grammatisk parsning och dessutom slutligen en felanalys kom tagguppsättningen att se ut på ett annat sätt än det gör i v. 2.0 av EPOS. Vissa syntaktiska särdrag skulle representeras för att möjliggöra grammatisk kontroll- och felanalys som rörde dessa. För t.ex. kongruens mellan adjektiv och nomen behövs sålunda numerus, genus, kasus och bestämdhet. En annan anledning till att tagguppsättningen såg annorlunda ut beror på att det inte hade konstruerats utifrån SMDB, utan byggde på version 11 av SAOL, vilken hade expanderats till alla möjliga morfologiska böjningsformer med hjälp av Sture Bergs böjningsgrupper och böjningsmönster; (se beskrivning av detta i Johansson Kokkinakis, 1999).

EPOS 1.0 följer EAGLES kriterier vad gäller obligatoriska attribut, i större utsträckning än EPOS 2.0. De rekommenderade attributen finns med i stor utsträckning; det finns dock ett antal taggar som anger kombinerade morfologiska egenskaper liknande dem i SUC, t.ex. genus=N/U, men dessa finns också som självständiga värden, t.ex. genus=N och genus=U. Liksom i EPOS 2.0 finns även här en del speciella extensioner.

Leechs kriterier angående kortfattade namn uppfylls optimalt (liksom i v. 2.0) då varje tagg i obligatoriska attribut representeras med endast en bokstav. Det andra kriteriet som gäller tydlighet följs delvis. Även här finns taggar som kan vara svåra att tolka, t.ex. "Q" för adverbialpartikel. Det tredje kriteriet som handlar om möjlighet att separera en tagg i dess beståndsdelar för att få fram vilka värden varje attribut innehåller, följs godtagbart på samma sätt som i v. 2.0.

Av Elworthys två kriterier tillgodoses det externa kriteriet med avseende på representation av särdrag som behövs i ett senare skede. Det interna kriteriet kan anses vara maximalt uppfyllt med tillgängliga resurser.

EPOS 1.0 och EPOS 2.0

V. 1.0 innehåller vissa kategorier som v. 2.0 inte innehåller. Dessa är t.ex. indelning av adverb i 14 antal undergrupper i stället för 4 som i v. 2.0, skillnad i markering av lexikaliska och modala hjälpverb samt markering av adverbialpartikel som ett obligatoriskt attribut. Skillnader i prestation mellan de olika tagguppsättningarna och slutsatser om relationen mellan

tagguppsättninginnehåll och prestation behandlas i kapitel 5.

Enligt Elworthys studie av variation i tagguppsättning borde precisionen öka, åtminstone vid disambiguering av kända homografer, då tagguppsättningen är större. Det är tveksamt om detta påstående kan anses vara språkoberoende. Ej heller kan påståendet generaliseras så att det gäller alla typer av förändringar i tagguppsättning. Viss morfologisk information, som t.ex. för verb distinktionen mellan lexikaliska verb och modala hjälpverb, bör öka precisionen vid disambiguering, liksom andra morfologiska indelningar vilka tydliggör ett ords användning i olika kontexter. Tufis (2000) och Megyesi (2001) har angett att prestationen blir bättre om en mindre tagguppsättning används vid taggning och utvärdering än vid träning. Det påståendet är mer logiskt och självklart än Elworthys. I vilket fall borde ett slags mått användas för att ange hur svårt det är, med avseende på en viss typ av kontext, att avgöra vilka värden som är rätt för diverse attribut, vid disambiguering av homografer. Måttet kunde innehålla information om tillgänglig syntaktisk och morfologisk information i en kontext. Dessutom är det nödvändigt att beskriva antal taggar, dvs. möjliga val, vid disambiguering av en homograf. Om ett sådant mått fanns skulle det underlätta bedömningen av hur bra precisionen egentligen är i förhållande till storleken eller innehållet i en tagguppsättning. Det skulle också ge en uppfattning om precision sett från ett språkoberoende perspektiv.

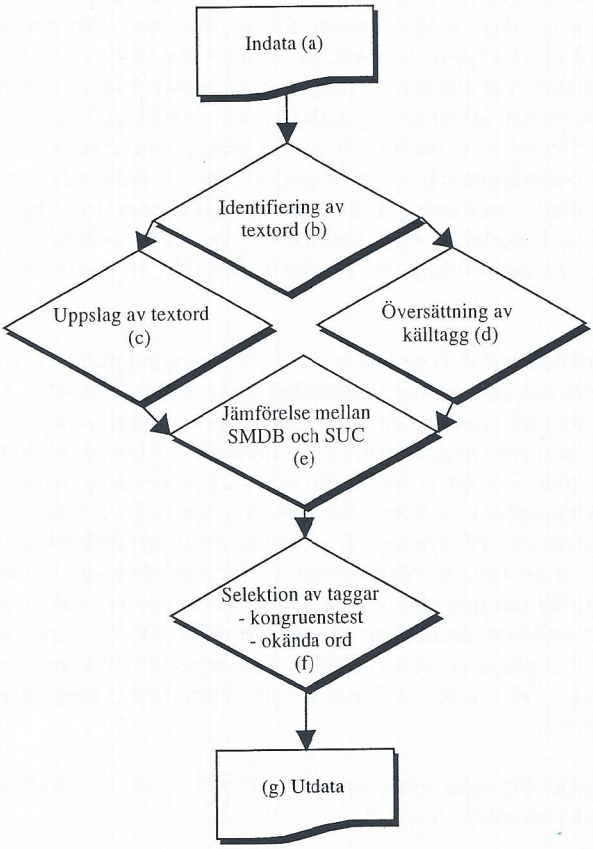
Sammanfattningsvis konstateras att de flesta tagguppsättningar har någon brist. Ofta kan det vara ej fullständigt representerade attribut i vissa ordklasser, som på grund av att vissa morfologiska beskrivningar inte har ansetts användbara i någon form av vidare analys. Krav på tydlighet förbises ofta, då man vanligtvis använder internationellt accepterade standarder, och då kan taggar t.ex. innehålla engelska termer. En slutsats man kan dra av ovanstående redogörelser för hur väl en tagguppsättning följer olika kriterier, är att oavsett vilket format på tagguppsättning man använder är det en fördel om inte alla men de flesta morfologiska beskrivningar finns med i en ordklass, åtminstone för samma språk. Det är annars svårt att utföra bearbetningar på material med olika tagguppsättningar, t.ex. vid s.k. mappning. För vidare diskussion om svårigheter med mappning, se avsnitt 4.1.3.

De tagguppsättningar vilka används i EPOS v. 1.0 och SMDB/EPOS v. 2.0, listas i appendix A och B.

4.1.2 Modulen "Suc2saol"

Den modul som har konstruerats för *mappning*, automatisk översättning, av SUC:s tagguppsättning till SMDB:s tagguppsättning i EPOS v. 2.0 benämns "suc2saol". Detta är en begränsad version av den modul som ingick i EPOS 1.0. Begränsningen, jämfört med tidigare version, består i att anropet av en subfunktion, vilken anger möjliga taggar i översättningen, inte ingår i version 2.0. Subfunktionen beskrivs i Berg et al. (2003).

Modulen som utför mappningen illustreras i figur 6. De ingående delarna i modulen anges generellt (a)-(g) nedan. Två kortare exempel ges i tabell 16 och tabell 17; ett längre exempel i tabell 18.



Figur 6. Modul för mappning av SUC:s tagguppsättning till SMDB:s tagguppsättning.

(a) *Indata* till mappningsmodulen består av tokeniserad och token-segmenterad text. Varje enskild teckensekvens anges på en särskild rad.

(b) *Identifiering av textord* består av att ett eller flera textord kopplas ihop till ett sammanhängande ord med en tillhörande tagg. Sekvenser som består av flera ord är t.ex. proprier. I exemplet (tabell 18) utgör "Don Kerr" ett sådant proprium. Dessa sekvenser av mer än ett ord är relativt lätta att urskilja eftersom de i SUC är taggade med en s.k. "NAME-tagg". Denna SGML-tagg används i SUC för att markera proprier av olika former. Informationen om proprier finns endast i SUC 2.0. NAME-taggen innehåller ett attribut "type" som beskriver typ av proprium. Det finns 9 typer och de översätts till någon av två motsvarande proprietaggar i SMDB:s tagguppsättning, "NPUSND" och "NPNSND" (tabell 15).

Kategori	SUC-info	SMDB-tagg	Exempel
NAME	type=PERSON	NPUSND	<w n=5>MATS<ana><ps>PM<m>NOM<b>Mats</w>
			<w n=6>LUNDEGÅRD<ana><ps>PM<m>NOM<b>Lundegård</w>
	type=PLACE	NPNSND	<w n=10>LONDON<ana><ps>PM<m>NOM<b>London</w>
	type=INST	NPNSND	<w n=7>DN:s<ana><ps>PM<m>GEN<b>DN</w>
	type=PRODUCT	NPNSND	<w n=46>Vit<ana><ps>JJ<m>POS UTR SIN IND NOM<b>vit</w>
			<w n=47>vintergylling<ana><ps>NN<m>UTR SIN IND NOM<b>vintergylling</w>
	type=WORK	NPNSND	<w n=1693>perestrojka<ana><ps>NN<m>UTR SIN IND NOM<b>perestrojka</w>
	type=EVENT	NPNSND	<w n=209>OS<ana><ps>PM<m>NOM<b>OS</w>
	type=OTHER	NPNSND	<w n=1476>Vita<ana><ps>JJ<m>POS UTR/NEU SIN DEF NOM<b>vit</w>
			<w n=1477>huset<ana><ps>NN<m>NEU SIN DEF NOM<b>hus</w>
	type=ANIMAL	NPUSND	<w n=1684>Rocky<ana><ps>PM<m>NOM<b>Rocky</w>
	type=MYTH	NPUSND	<w n=628>Vår<ana><ps>PS<m>UTR SIN DEF<b>vår</w>
			<w n=629>Herres<ana><ps>PM<m>GEN<b>Herre</w>

Tabell 15. Kategorisering av proprier i SUC, Källgren (1998:31).

c) *Uppslag av textord* innebär att en "lookup-funktion" anropas, nämligen subfunktionen "lookup_saol", skapad av Cederholm (Berg et al., 2003). Resultatet av funktionsanropet är att alla möjliga taggar för textordet returneras. Denna operation är nödvändig för att göra en jämförelse mellan de taggar som angivits i SMDB för ett ord och den/de tagg/taggar som är ett resultat från mappningsfunktionen, där det ev. kan förekomma skillnader. Skillnaderna består, t.ex. för ordet "andra", i att den mappade versionen av SUC:s tagg "JJ POS UTR/NEU PLU IND/DEF NOM" blir "AQP*PN*". Om man slår upp "andra" i SAOL/SMDB, så har ordet sju möjliga taggar, men ingen utav dessa är en adjektivtagg. Ett exempel med "andra" finns i tabell 16. I det angivna exemplet bör SMDB:s "DI*P-tagg" användas, men en analys av vilken tagg som är rätt i vilket sammanhang kan ännu inte göras automatiskt, med undantag för kontexter där särskilda ord kan specificeras. Därför kan sådana problemfall inte lösas idag utan kräver manuell analys. Detta har utförts i 10% av SUC, dvs. den del som använts i utvärderingen.

Det finns några undantag, som har studerats och numera åtgärdas inom disambigueringsmodulen. T.ex. ordet "själv", taggas med SMDB som "PFUS" i de fall då det i SUC är taggat som adjektiv, ordet "mycket", taggas med SMDB som "AQPNSNI" i de fall då det taggas i SUC som pronomen och "liksom" taggas i SMDB som "CC" i de fall då det i SUC taggas som preposition.

(d) *Översättning av källtagg* innebär att en SUC-tagg ska översättas till en SMDB-tagg. Operationen utförs satsvis, ett textord åt gången, genom att anropa en subfunktion "map_tagset", skapad av Cederholm (Berg et al., 2003). Resultatet av funktionsanropet är att ett antal möjliga taggar i SMDB-format returneras. Optimalt returneras endast en tagg, men ofta finns det inte ett 1:1-förhållande mellan tagguppsättningarna och då returneras flera taggar (se exemplet i tabell 17). Där kan man se att taggen "JJ POS UTR/NEU SIN DEF NOM" i SUC får motsvarigheten "AQP*SND" vid mappning. Vid närmare undersökning av vilka taggar som finns i SMDB, framgår det att taggen inte finns i exakt motsvarighet utan motsvaras i stället av två taggar, "AQPUSND_kompromisslös.1|AQPNSND_kompromisslös.1". Asterisksymbolen anger att vilket som helst av alla värden på attributet kan användas. Då krävs vidare analys för att avgöra vilken tagg som är rätt. Till skillnad från det tidigare exemplet då det inte fanns motsvarande taggar i SMDB, får man här följande svar om man slår upp "kompromisslösa" i SMDB: "kompromisslösa/[AQPUSND_kompromisslös.1|AQPNSND_kompromisslös.1|AQP*PN*_kompromisslös.1]". Detta betyder att det finns flera taggar som är gemensamma i mängderna av taggar från den mappade taggen och taggarna från SMDB. Denna operation utförs i nästa steg (e).

e) *Jämförelse mellan SMDB och SUC* sker mellan resultaten från översättningen av SUC-tagg och resultatet från "lookup-funktionen" för ett ord. Finns det en gemensam tagg är det denna som väljs som den slutgiltiga taggen i mappningen, men finns det flera gemensamma går analysen vidare till nästa steg (f). Finns det inga gemensamma taggar rapporteras ett fel.

f) *Selektion av möjliga taggar* är en analys som tillämpas då resultat från tidigare analyser resulterat i mer än en resulterande tagg för ett textord. Målet med denna analys är att om möjligt begränsa taggarna till endast en. Här utförs kongruenstest om det är så att t.ex. en tagg i SUC resulterar i två taggar i SMDB. (Se exemplet i tabell 18 med "det ansedda analysinstitutet"). Då tillämpas en analys som automatiskt, efter tillgänglig kongruensinformation i kontexten, väljer rätt tagg. I exemplet finns information om genus och numerus i taggen före, neutrum och singular, och då används den informationen för att välja rätt tagg, "AF0NSND". Då okända ord uppträder i texten, får de den tagg som anges i mappningen.

g) *Utdata*. Där har i bästa fall varje textord endast en tagg, ibland flera då korrekt taggvariant inte har kunnat identifieras eller disambiguerats. Disambiguering kan dessbättre utföras manuellt i efterhand. De övriga problemen består. Vad gäller utdata från mappningsmodulen bör nämnas att innehållet i textmaterialet i princip är detsamma i utdata som indata från SUC-materialet, förutom att det inte finns med någon information om ordens grundformer. Eftersom syftet med mappningen är att skapa utgångsmaterial för träning av en taggare behövs grundformerna inte för träningen.

Del	Uppgift i modulen	Exempel
a	Indata	<w n=118>på<ana><ps>PP<b>på</w> <w n=119>andra<ana><ps>JJ<m>POS UTR/NEU PLU IND/DEF NOM<b>annan</w> <w n=120>håll<ana><ps>NN<m>NEU PLU IND NOM<b>håll</w>
b	Identifiering av textord	på PP andra JJ<m>POS UTR/NEU PLU IND/DEF NOM håll NN<m>NEU PLU IND NOM
c	Uppslag av textord	på [R0_på.2|S_på.1] andra [MO*SND_andra.1|V0N0A_andra.2|V0M0A_andra.2| DI*S_annan.1|PI*S*_annan.1|DI*P_annan.1| PI*P*_annan.1] håll [NCNSNI_håll.1|NCNPNI_håll.1|V0M0A_hålla.1]
d	Översättning av källtagg	på PP --> S andra JJ POS UTR/NEU PLU IND/DEF NOM --> AQP*PN* håll NN NEU PLU IND NOM --> NCNPNI

Tabell 16. Exemplet "på andra håll" från taggning med SMDB och EPOS.

Del	Uppgift i modulen	Exempel
e	Jämförelse mellan SMDB och SUC	på S <=> S_på.1 andra AQP*PN* <=> [MO*SND_andra.1|V0N0A_andra.2| V0M0A_andra.2|DI*S_annan.1|PI*S*_annan.1| DI*P_annan.1|PI*P*_annan.1] håll NCNPNI <=> NCNPNI_håll.1
f	Selektion av taggar - kongruenstest - okända ord	på S andra AQP*PN* <=> ? (finns ingen motsvarighet i SMDB) håll NCNPNI
g	Utdata	på S andra ? (finns ingen motsvarighet i SMDB) håll NCNPNI

Tabell 16. Exemplet "på andra håll" från taggning med SMDB och EPOS.

Del	Uppgift i modulen	Exempel
a	Indata	<w n=871>att<ana><ps>SN<b>att</w> <NAME TYPE=PERSON> <w n=872> Landsbergis<ana> <ps> PM<m> GEN<b> Landsbergis</w> </NAME> <w n=873>kompromisslösa<ana><ps>JJ<m>POS UTR/NEU SIN DEF NOM<b>kompromisslös</w> <d n=874>,<ana><ps>MID<b>,</d> <w n=875>stundom<ana><ps>AB<b>stundom</w> <w n=876>utmanande<ana><ps>PC<m>PRS UTR/NEU SIN/PLU IND/DEF NOM<b>utmanande</w> <d n=877>,<ana><ps>MID<b>,</d>
b	Identifiering av textord	att SN <NAME TYPE=PERSON> Landsbergis PM<m>GEN </NAME> kompromisslösa JJ<m>POS UTR/NEU SIN DEF NOM , MID stundom AB utmanande PRS UTR/NEU SIN/PLU IND/DEF NOM , MID
c	Uppslag av textord	att/[Z_att.2|CS_att.1] Landsbergis [] kompromisslösa/ [AQPUSND_kompromisslös.1|AQPNSND_kompromisslös.1|AQP*PN*_kompromisslös.1] , [FI] stundom/[R0_stundom.1] utmanande/ [AP0**N*_utmana.1|R0_utmana.1|NCNSNI_utmana.1|AQPUSNI_utmanande.1|AQPNSNI_utmanande.1|AQPUSND_utmanande.1|AQPNSND_utmanande.1|AQP*PN*_utmanande.1|R0_utmanande.1] , [FI]

Tabell 17. Exemplet " ... att Landsbergis kompromisslösa, stundom utmanande, ...", från taggning med SMDB och EPOS.

Del	Uppgift i modulen	Exempel
d	Översättning av källtagg	att SN --> CS Landsbergis <NAME TYPE=PERSON> PM GEN--> NPNSGD kompromisslösa JJ POS UTR/NEU SIN DEF NOM-->AQP*SND , MID --> FI stundom AB-->R0 utmanande PC PRS UTR/NEU SIN/PLU IND/DEF NOM-->AP0**N* , MID --> FI
e	Jämförelse mellan SMDB och SUC	att CS <=> CS_att.1 Landsbergis NPNSGD kompromisslösa AQP*SND <=> [AQPUSND_kompromisslös.1| AQPNSND_kompromisslös.1] , FI stundom R0 <=> R0_stundom.1 utmanande AP0**N* <=> AP0**N*_utmana.1 , FI
f	Selektion av taggar - kongruenstest - okända ord	att CS Landsbergis NPNSGD kompromisslösa AQPUSND el. AQPNSND (ingen info i närkontexten) , FI stundom R0 utmanande AP0**N* , FI
g	Utdata	att CS Landsbergis NPNSGD kompromisslösa AQPUSND,AQPNSND , FI stundom R0 utmanande AP0**N* , FI

Tabell 17. Exemplet " ... att Landsbergis kompromisslösa, stundom utmanande, ...", från taggning med SMDB och EPOS.

Del	Uppgift i modulen	Exempel
a	Indata	<s id=aa01a-006><NAME TYPE=PERSON> <w n=44>Don<ana><ps>PM<m>NOM<b>Don</w> <w n=45>Kerr<ana><ps>PM<m>NOM<b>Kerr</w> </NAME> <d n=46>,<ana><ps>MID<b>,</d> <w n=47>en<ana><ps>PN<m>UTR SIN IND SUB/OBJ<b>en</w> <w n=48>av<ana><ps>PP<b>av</w> <w n=49>de<ana><ps>DT<m>UTR/NEU PLU DEF<b>den</w> <w n=50>politiska<ana><ps>JJ<m>POS UTR/NEU PLU IND/DEF NOM<b>politisk</w> <w n=51>tänkarna<ana><ps>NN<m>UTR PLU DEF NOM<b>tänkare</w> <w n=52>på<ana><ps>PP<b>på</w> <w n=53>det<ana><ps>DT<m>NEU SIN DEF<b>den</w> <w n=54>ansedda<ana><ps>PC<m>PRF UTR/NEU SIN DEF NOM<b>ansedd</w> <w n=55>analysinstitutet<ana><ps>NN<m>NEU SIN DEF NOM<b>analysinstitut</w> <NAME TYPE=INST><ABBR> <w n=56>IISS<ana><ps>PM<m>NOM<b>IISS</w> </ABBR></NAME> <d n=57>,<ana><ps>MID<b>,</d> <w n=58>är<ana><ps>VB<m>PRS AKT<b>vara</w> <w n=59>inte<ana><ps>AB<b>inte</w> <w n=60>påfallande<ana><ps>AB<b>påfallande</w> <w n=61>optimistisk<ana><ps>JJ<m>POS UTR SIN IND NOM<b>optimistisk</w> <w n=62>när<ana><ps>HA<b>när</w> <w n=63>han<ana><ps>PN<m>UTR SIN DEF SUB<b>han</w> <w n=64>talar<ana><ps>VB<m>PRS AKT<b>tala</w> <w n=65>om<ana><ps>PP<b>om</w> <w n=66>saken<ana><ps>NN<m>UTR SIN DEF NOM<b>sak</w> <d n=67>.<ana><ps>MAD<b>.</d></s>

Tabell 18. Exempel "Don Kerr, en av de politiska tänkarna ... ".

Del	Uppgift i modulen	Exempel
b	Identifiering av textord	<NAME TYPE=PERSON>Don PM<m>NOM Kerr PM<m>NOM</NAME> , MID en PN<m>UTR SIN IND SUB/OBJ av PP de DT<m>UTR/NEU PLU DEF politiska JJ<m>POS UTR/NEU PLU IND/DEF NOM tänkarna NN<m>UTR PLU DEF NOM på PP det DT<m>NEU SIN DEF ansedda PC<m>PRF UTR/NEU SIN DEF NOM analysinstitutet NN<m>NEU SIN DEF NOM <NAME TYPE=INST>IISS PM<m>NOM</NAME> , MID är VB<m>PRS AKT inte AB påfallande AB optimistisk JJ<m>POS UTR SIN IND NOM när HA han PN<m>UTR SIN DEF SUB talar VB<m>PRS AKT om PP saken NN<m>UTR SIN DEF NOM . MAD

Tabell 18. Exempel "Don Kerr, en av de politiska tänkarna ... ".

Del	Uppgift i modulen	Exempel
c	Uppslag av textord	Don Kerr [] en [NCUSNI_en.5|R0_en.4|DIUS_en.2|MCUSNI_en.1|PIUS*_en.3] av [S_av.1|R0_av.2] de [DF*P_den.2|PF*PS_den.1] politiska [AQPUSND_politisk.1|AQPNSND_politisk.1|AQP*PN*_politisk.1] tänkarna [NCUPND_tänkare.1] på [R0_på.2|S_på.1] det [DFNS_den.2|PFNS*_den.1] ansedda [AF0USND_anse.1|AF0NSND_anse.1|AF0*PN*_anse.1|AQPUSND_ansedd.1|AQPNSND_ansedd.1|AQP*PN*_ansedd.1] analysinstitutet [] IISS [] , [FI] är [V0IPA_vara.4] inte [R0_inte.1] påfallande [AQPUSNI_påfallande.1|AQPNSNI_påfallande.1|AQPUSND_påfallande.1|AQPNSND_påfallande.1|AQP*PN*_påfallande.1|R0_påfallande.1] optimistisk [AQPUSNI_optimistisk.1] när [CS_när.2|R0_när.1|V0IPA_nära.3|V0M0A_nära.3] han [PFUSS_han.1] talar [V0IPA_tala.1|NCUSNI_talar.1] om [CS_om.1|S_om.2|R0_om.2] saken [NCUSND_sak.1] . [FE]

Tabell 18. Exempel "Don Kerr, en av de politiska tänkarna ... ".

Del	Uppgift i modulen	Exempel
d	Översättning av källtagg	Don Kerr <NAME TYPE=PERSON></NAME> --> NPUSND , MID -->FI en PN UTR SIN IND SUB/OBJ-->PIUS* av PP-->S de DT UTR/NEU PLU DEF-->DF*P politiska JJ POS UTR/NEU PLU IND/DEF NOM-->AQP*PN* tänkarna NN UTR PLU DEF NOM-->NCUPND på PP-->S det DT NEU SIN DEF-->DFNS ansedda PC PRF UTR/NEU SIN DEF NOM-->AF0*SND analysinstitutet NN NEU SIN DEF NOM-->NCNSND IISS <NAME TYPE=INSTITUTE></NAME> --> NPNSND , MID-->FI är VB PRS AKT-->V0IPA inte AB-->R0 påfallande AB-->R0 optimistisk JJ POS UTR SIN IND NOM-->AQPUSNI när HA-->R0 han PN UTR SIN DEF SUB-->PFUSS talar VB PRS AKT-->V0IPA om PP-->S saken NN UTR SIN DEF NOM-->NCUSND . MAD-->FE
e	Jämförelse mellan SMDB och SUC	Don Kerr NPUSND , FI en PIUS* <=> PIUS*_en.3 av S <=> S_av.1 de DF*P <=> DF*P_den.2 politiska AQP*PN* <=> AQP*PN*_politisk.1 tänkarna NCUPND <=> NCUPND_tänkare.1 på S <=> S_på.1 det DFNS <=> DFNS_den.2 ansedda AF0*SND <=> [AF0USND_anse.1|AF0NSND_anse.1|AF0*PN*_anse.1|AQPUSND_ansedd.1|AQPNSND_ansedd.1|AQP*PN*_ansedd.1] analysinstitutet NCNSND <=> [] IISS NPNSND , FI är V0IPA <=> V0IPA_vara.4 inte R0 <=> R0_inte.1 påfallande R0 <=> R0_påfallande.1 optimistisk AQPUSNI <=> AQPUSNI_optimistisk.1 när R0 <=> R0_när.1 han PFUSS <=> PFUSS_han.1 talar V0IPA <=> V0IPA_tala.1 om S <=> S_om.2 saken NCUSND <=> NCUSND_sak.1 . FE

Tabell 18. Exempel "Don Kerr, en av de politiska tänkarna ... ".

Del	Uppgift i modulen	Exempel
f	Selektion av taggar - kongruenstest - okända ord	Don Kerr NPUSND , FI en PIUS* av S de DF*P politiska AQP*PN* tänkarna NCUPND på S det DFNS ansedda [AF0USND_anse.1|AF0NSND_anse.1] -> AF0NSND *kongruenstest* analysinstitutet NCNSND *okänt ord* IISS NPNSND , FI är V0IPA inte R0 påfallande R0 optimistisk AQPUSNI när R0 han PFUSS talar V0IPA om S saken NCUSND . FE
g	Utdata	Don Kerr NPUSND , FI en PIUS* av S de DF*P politiska AQP*PN* tänkarna NCUPND på S det DFNS ansedda AF0NSND analysinstitutet NCNSND IISS NPNSND , FI är V0IPA inte R0 påfallande R0 optimistisk AQPUSNI när R0 han PFUSS talar V0IPA om S saken NCUSND . FE

Tabell 18. Exempel "Don Kerr, en av de politiska tänkarna ... ".

4.1.3 Skillnader och problem

I detta avsnitt redogörs för några av de skillnader som finns mellan tagguppsättningarna i SUC och SMDB samt de problem som uppstod vid mappningen.

Här ges en redovisning av skillnader i tagguppsättningarna ordklassvis, tillsammans med slutsatser om vilken betydelse skillnaderna kan ha.

Problem i mappning kan uppdelas i två huvudgrupper: dels problem med koherens i översättningen mellan tagguppsättningar i mappningen, dels skillnader i taggning, t.ex. subjektiva val av taggar, olika lexikon som utgångspunkt samt feltaggning.

4.1.3.1 Skillnader i tagguppsättningar

SUC:s tagguppsättning och SMDB:s tagguppsättning, som redovisades kortfattat i föregående avsnitt, består av 153 respektive 133 taggar. Förutom de uppenbara syntaktiska/morfologiska skillnaderna i namn, skiljer sig tagguppsättningarna åt på vissa punkter i följande ordklasser; Bestämd artikel, räkneord, nomen, egennamn, pronomen, adverb, verb och förkortningar. I tabell 19 anges fördelningen mellan tagguppsättningarna samt antalet subklasser inom ordklasserna. I tabell 20–tabell 27 ges, för varje ordklass förklaringar av skillnader i taggar, med exempel och kommentarer. Existerande skillnader i ordklasser kommenteras i det följande.

I tabell 19 redovisas tagguppsättningen i EPOS 1.0 som har 282 taggar. Dessa kommer inte att jämföras med de övriga tagguppsättningarna annat än i slutet av avsnittet p.g.a. att EPOS 1.0 inte används längre. I utvärderingen dras slutsatser om skillnader i prestation i förhållande till storleken på tagguppsättningarna, även med EPOS 1.0.

Ordklass	SUC-taggar	SMDB-taggar	EPOS 1.0
Adjektiv	37	37	97
Konjunktioner	2	2	7
Best. artikel	17	16	7
Interpunktion	3	3	1
Interjektion	1	1	1
Räkneord	10	9	2
Nomen	22	19	88

Tabell 19. Fördelning av taggar i SUC:s och SMDB:s tagguppsättningar.

Ordklass	SUC-taggar	SMDB-taggar	EPOS 1.0
Egennamn	3	2	3
Pronomen	25	24	20
Adverb	7	4	14
Preposition	1	1	1
Verb	14	12	38
Utländska ord	1	1	1
Infinitivmärke	1	1	1
Förkortning	9	1	1
Summa	153	133	282

Tabell 19. Fördelning av taggar i SUC:s och SMDB:s tagguppsättningar.

Ordklass	SMDB-tagg	SUC-taggar
BESTÄMD ARTIKEL	DF*P	DT UTR/NEU PLU DEF, DT UTR/NEU PLU IND/DEF
Exempel		
SUC (ex. 1)	<w n=357>Alla<ana><ps>DT<m>UTR/NEU PLU IND/DEF<b>all</w>	
	<w n=358>sorter<ana><ps>NN<m>UTR PLU IND NOM<b>sort</w>	
SMDB	Alla/[ERROR_DIFF:DF*P-NCUSNI,D0*P,PI*P*] sorter/NCUPNI	
Förklaring	I exemplet ovan ska "Alla" taggas som artikel. I gruppen artikel har SMDB ingen generell form som betecknar både bestämd och obestämd form för att motsvara SUC:s tagg "DT UTR/NEU PLU IND/DEF", DF*P i mappat format. SMDB har däremot motsvarigheter för de mer specifika formerna "DT UTR/NEU PLU DEF" och "DT UTR/NEU PLU IND". "Alla" i SMDB taggas med någon av följande taggar "NCUSNI,D0*P,PI*P*", därmed får man en konflikt i mappningen. I SUC har 1175 fall av "alla" taggats med denna tagg.	

Tabell 20. Skillnader vid artikel, exempel från mappningsarbetet mellan SUC och SMDB.

Ordklass	SMDB-tagg	SUC-taggar
RÄKNEORD	MC*PN*	RG NOM, RG SMS

Tabell 21. Skillnader vid räkneord, exempel från mappningsarbetet mellan SUC och SMDB.

Ordklass	**SMDB-tagg**	**SUC-taggar**
Exempel		
SUC (ex. 2)	<num><w n=398>två<ana><ps>RG<m>NOM<b>två</w></num>	
	<w n=399>nya<ana><ps>JJ<m>POS UTR/NEU PLU IND/DEF NOM<b>ny</w>	
	<w n=400>missiler<ana><ps>NN<m>UTR PLU IND NOM<b>mis-sil</w>	
SMDB	två/MC*PN* nya/AQP*PN* missiler/NCUPNI	
SUC (ex. 3)	<num><w n=2168>1950-<ana><ps>RG<m>SMS<b>1950</w></num>	
	<w n=2169>och<ana><ps>KN<b>och</w>	
	<w n=2170>1960-talen<ana><ps>NN<m>NEU PLU DEF NOM<b>1960-tal</w>	
SMDB	1950-/MC*PN* och/CC 1960-talen/NCNPND	
Förklaring	SMDB:s tagg MC*PN motsvaras av SUC:s taggar RG NOM och RG SMS vilka används för att tagga två olika typer av ord, nämligen vanliga räkneord, se ex 2 ovan och sammansatta årtalsuttryck, se ex 3 ovan. I detta fall har SUC en mer specifik taggning än vad som är möjligt med SMDB.	

Tabell 21. Skillnader vid räkneord, exempel från mappningsarbetet mellan SUC och SMDB.

Skillnader i nomen uppkommer bl. a. därför att SUC ofta skiljer ut former som är någon slags sammansättning, vilket SMDB inte markerar; därför blir SMDB:s taggar ibland samlingstaggar för mer än en annan SUC-tagg då det handlar om t.ex. sammansättningar. Detta gäller också egennamn, pronomen och i viss mån adverb.

Ordklass	**SMDB-tagg**	**SUC-taggar**
NOMEN	NC0000	NN - - - SMS, NN - - - -
Exempel		
SUC (ex. 4)	<w n=2287>styr-<ana><ps>NN<m>- - - SMS<b>styr</w>	
	<w n=2288>och<ana><ps>KN<b>och</w>	
	<w n=2289>reglersystem<ana><ps>NN<m>NEU PLU IND NOM<b>reglersystem</w>	
SMDB	styr-/NC*000 och/CC reglersystem/NCNPNI	
SUC (ex. 5)	<w n=388>i<ana><ps>PP<b>i</w>	
	<w n=389>fjol<ana><ps>NN<m>- - - -<b>fjol</w>	

Tabell 22. Skillnader vid nomen, exempel från mappningsarbetet mellan SUC och SMDB.

Ordklass	SMDB-tagg	SUC-taggar
SMDB	i/S fjol/NC0000	
Förklaring	SMDB:s tagg NC0000 motsvaras av SUC:s taggar NN - - - SMS och NN - - - - vilka används för att tagga två olika typer av ord, nämligen nomen som ingår i t.ex. adverb eg. "i fjol", se ex 5 ovan samt sammansatta nomen, se ex 4 ovan. I detta fall har SUC en mer specifik taggning än vad som är möjligt med SMDB.	

Tabell 22. Skillnader vid nomen, exempel från mappningsarbetet mellan SUC och SMDB.

Ordklass	SMDB-tagg	SUC-taggar
EGENNAMN	NP*SND	PM NOM, PM SMS
Exempel		
SUC (ex. 6)	<NAME TYPE=PERSON>	
	<w n=5>MATS<ana><ps>PM<m>NOM<b>Mats</w>	
	<w n=6>LUNDEGÅRD<ana><ps>PM<m>NOM<b>Lundegård</w></NAME>	
SMDB	MATS LUNDEGÅRD/NPUSND	
SUC (ex. 7)	<NAME TYPE=PLACE>	
	<w n=1866>Göteborgs-<ana><ps>PM<m>SMS<b>Göteborg</w>	
	</NAME>	
	<w n=1867>och<ana><ps>KN<b>och</w>	
	<NAME TYPE=PLACE>	
	<w n=1868>Bohuslän<ana><ps>PM<m>NOM<b>Bohuslän</w></NAME>	
SMDB	Göteborgs-/NPNSND och/CC Bohuslän/NPNSND	
Förklaring	SMDB:s tagg NP*SND motsvaras av SUC:s taggar PM NOM och PM SMS vilka används för att tagga två olika typer av ord, nämligen vanliga egennamn, se ex 6 ovan, samt sammansatta egennamn, se ex 7 ovan. I detta fall har SUC en mer specifik taggning än vad som är möjligt med SMDB.	

Tabell 23. Skillnader vid egennamn, exempel från mappningsarbetet mellan SUC och SMDB.

Ordklass	**SMDB-tagg**	**SUC-taggar**
PRONOMEN	PHNS0	HP NEU SIN IND, HP NEU SIN IND SMS
Exempel		
SUC (ex. 8)	<w n=115>vad<ana><ps>HP<m>NEU SIN IND<b>vad</w>	
	<w n=116>som<ana><ps>HP<m>- - -<b>som</w>	
	<w n=117>händer<ana><ps>VB<m>PRS AKT<b>hända</w>	
SMDB	vad/PHNS0 som/PH**0 händer/V0IPA	
SUC (ex. 9)	<w n=598>vad-<ana><ps>HP<m>NEU SIN IND SMS<b>vad</w>	
	<w n=599>och<ana><ps>KN<b>och</w>	
	<w n=600>hur-frågorna<ana><ps>NN<m>UTR PLU DEF NOM<b>hur-fråga</w>	
SMDB	vad-/PHNS0 och/CC hur-frågorna/NCUPND	
Förklaring	SMDB:s tagg PHNS0 motsvaras av SUC:s taggar HP NEU SIN IND och HP NEU SIN IND SMS vilka används för att tagga två olika typer av ord, nämligen vanliga frågande eller relativa pronomen, se ex 8 ovan, samt sammansatta relativa pronomen, se ex 9 ovan. I detta fall har SUC en mer specifik taggning än vad som är möjligt med SMDB.	

Tabell 24. Skillnader vid pronomen, exempel från mappningsarbetet mellan SUC och SMDB.

Ordklass	**SMDB-tagg**	**SUC-taggar**
ADVERB	R0	AB, HA, PL, AB SMS
Exempel		
SUC (ex. 10)	<w n=20>inte<ana><ps>AB<b>inte</w>	
	<w n=21>mycket<ana><ps>JJ<m>POS NEU SIN IND NOM<b>mycken</w>	
SMDB	inte/R0 mycket/AQPNSNI	
SUC (ex. 11)	<w n=70>vad<ana><ps>HA<b>vad</w>	
	<w n=71>gäller<ana><ps>VB<m>PRS AKT<b>gälla</w>	
	<w n=72>missiler<ana><ps>NN<m>UTR PLU IND NOM<b>missil</w>	
SMDB	vad/R0 gäller/V0IPA missiler/NCUPNI	

Tabell 25. Skillnader vid adverb, exempel från mappningsarbetet mellan SUC och SMDB.

Ordklass	SMDB-tagg	SUC-taggar
SUC (ex. 12)	<w n=188>bygger<ana><ps>VB<m>PRS AKT<b>bygga</w>	
	<w n=189>under<ana><ps>PL<b>under</w>	
	<w n=190>spänningarna<ana><ps>NN<m>UTR PLU DEF NOM<b>spänning</w>	
SMDB	bygger/V0IPA under/R0 spänningarna/NCUPND	
SUC (ex. 13)	<w n=892>upp-<ana><ps>AB<m>SMS<b>upp</w>	
	<w n=893>och<ana><ps>KN<b>och</w>	
	<w n=894>nerflaggningar<ana><ps>NN<m>UTR PLU IND NOM<b>nerflaggning</w>	
SMDB	upp-/R0 och/CC nerflaggningar/NCUPNI	
Förklaring	Exemplen ovan är användningar av SUC:s taggar AB, HA, PL samt AB SMS vilka alla taggas med en tagg i SMDB, nämligen R0. AB och HA hör till funktionsord i SUC. Partikel finns inte i SMDB utan representeras som adverb. Sammansättningar med adverb markeras liksom andra sammansättningar inte med SMDB. Se exempel 10, 11, 12 och 13. I alla dessa fall är taggningen mer specifik i SUC än i SMDB.	

Tabell 25. Skillnader vid adverb, exempel från mappningsarbetet mellan SUC och SMDB.

I SUC anges specifik information i verb om konjunktivformer, vilket inte görs i SMDB. Dessutom förekommer flera exempel på talspråk i SUC, vilket inte representeras i SMDB. Konsekvensen blir att ord av denna typ, t.ex. preteritumformer i stället tolkas som infinitiv- eller imperativformer.

Ordklass	SMDB-tagg	SUC-taggar
VERB	V0SPA, V0ISA	VB KON PRS AKT, VB PRS AKT
Exempel		
SUC (ex. 14)	<w n=2485>Må<ana><ps>VB<m>KON PRS AKT<b>må</w>	
	<w n=2486>biskopen<ana><ps>NN<m>UTR SIN DEF NOM<b>biskop</w>	
	<w n=2487>med<ana><ps>PP<b>med</w>	
	<NAME TYPE=MYTH>	
	<w n=2488>Guds<ana><ps>PM<m>GEN<b>Gud</w>	
	</NAME>	
	<w n=2489>hjälp<ana><ps>NN<m>UTR SIN IND NOM<b>hjälp</w>	

Tabell 26. Skillnader vid verb, exempel från mappningsarbetet mellan SUC och SMDB.

Ordklass	SMDB-tagg	SUC-taggar
SMDB	Må/[ERROR_DIFF:V0SPA-V0N0A,V0M0A,R0,V0S**] biskopen/ NCUSND med/S Guds/NPUSGD hjälp/NCUSNI	
SUC (ex. 15)	<w n=272>eleverna<ana><ps>NN<m>UTR PLU DEF NOM<b>elev</ w>	
	<d n=273>)<ana><ps>PAD<b>)</d>	
	<w n=274>kasta<ana><ps>VB<m>PRT AKT<b>kasta</w>	
	<w n=275>snöboll<ana><ps>NN<m>UTR SIN IND NOM<b>snöboll</w>	
	<w n=276>på<ana><ps>PP<b>på</w>	
	<w n=277>varann<ana><ps>PN<m>UTR/NEU PLU DEF OBJ<b>varandra</w>	
SMDB	eleverna/NCUPND)/FC kasta/[ERROR_DIFF:V0ISA-V0N0A,V0M0A] snöboll/NCUSNI på/S varann/PFUP0	
Förklaring	Exempel 14 ovan anger ett ord taggat som konjunktiv, vilket är en form som kan representeras med SMDB:s tagguppsättning, men där taggen till ordet inte beskriver tempus el. liknande i SMDB. De ord som har konjunktivformer i SMDB kan bara få en tagg och det är V0S**, vilket är en mer generell form än motsvarande taggar för konjunktiv i SUC. Exempel 15 ovan, är ett exempel på talspråk i SUC. Eftersom talspråk med sk. 'eliderade' former, som t.ex. 'kasta' i stället för 'kastade' inte finns med i SMDB i formen preteritum, utan endast infinitiv och imperativ, blir det en konflikt. Det finns många exempel på talspråk i SUC, och det är inte självklart hur dessa fall ska behandlas. Eftersom det inte rör sig om skrivfel bör man inte korrigera orden i korpusen, men det finns heller ingen översättning av taggar som fungerar tillfredsställande. Problemfallen kvarstår alltså.	

Tabell 26. Skillnader vid verb, exempel från mappningsarbetet mellan SUC och SMDB.

I SUC markeras förkortningar inom en ordklass. I SMDB markeras inte förkortningar i enskilda ordklasser utan endast en. Därför blir SMDB:s förkortningstagg en samlingstagg för nio andra SUC-taggar vilka beskriver förkortningar inom olika ordklasser.

Ordklass	SMDB-tagg	SUC-taggar
FÖRKORTNING	Y	JJ AN, AB AN, PP AN

Tabell 27. Skillnader vid förkortningar, exempel från mappningsarbetet mellan SUC och SMDB.

Ordklass	SMDB-tagg	SUC-taggar
Exempel		
SUC (ex. 16)	<ABBR>	
	<w n=800>höggrefl.<ana><ps>JJ<m>AN<b>höggrefl</w></ABBR>	
	<ABBR>	
	<w n=801>ex:s<ana><ps>NN<m>AN<b>ex:s</w></ABBR>	
SMDB	höggrefl./Y ex:s/Y	
SUC (ex. 17)	<ABBR>	
	<w n=69>kl<ana><ps>AB<m>AN<b>kl</w></ABBR>	
	<num>	
	<w n=70>11<ana><ps>RG<m>NOM<b>11</w></num>	
SMDB	kl/Y 11/MC*PN*	
SUC (ex. 18)	<ABBR>	
	<w n=90>inkl.<ana><ps>PP<m>AN<b>inkl</w></ABBR>	
	<w n=91>stat<ana><ps>NN<m>UTR SIN IND NOM<b>stat</w>	
SMDB	inkl./Y stat/NCUSNI	
Förklaring	Förkortningar taggas i SMDB med Y, och i SUC finns det nio olika förkortningstaggar; PS AN, AB AN, KN AN, DT AN, NN AN, PP AN, VB AN, JJ AN, PC AN, några exempel visas visas, ex. 14, 15, 16.	

Tabell 27. Skillnader vid förkortningar, exempel från mappningsarbetet mellan SUC och SMDB.

Jämfört med EPOS 2.0 och SMDB innehåller SUC:s tagguppsättning fler morfologiska distinktioner. Hur en mindre tagguppsättning påverkar prestationen av en taggare är svårt att bedöma på grund av att det inte endast är storleken på tagguppsättningen som påverkar prestationen, utan också vilken morfologisk information som finns representerad. Andra faktorer som spelar in är vilken information som finns om homografer i lexikonet som används och vilken metod taggaren använder sig av vid disambiguering av homografer. Konkretisering av ett exempel där SMDB inte representerar information som finns i SUC:s tagguppsättning, är de adverb som kallas adverbialpartiklar, t.ex. "gå ut", där "ut" är en sådan. Detta innebär att informationen om dessa partiklar försvinner då de i mappningsprocessen översätts till enbart adverb, "RO". Vid träning av taggaren kommer det därför att bildas fler regler vilka innehåller "RO"-taggen än om det också fanns en "PL"-tagg, dvs. partikel. Konsekvenserna som följer när en taggare ska disambiguera homografer, är att adverbialpartikel aldrig kan identifieras. Det blir färre fel då val mellan homograferna "RO" och "PL" undviks. Disambiguering av intern homografi inom ett adverb för-

ändras inte av detta.

Vid jämförelse mellan EPOS 1.0, som bygger på SAOL11, och SUC:s tagguppsättning, innehåller det senare färre morfologiska distinktioner i ordklasserna adjektiv, konjunktioner, nomen, adverb och verb. I klasserna bestämd artikel, interpunktion, räkneord, pronomen och förkortning finns fler morfologiska distinktioner i SUC. I interjektion, egennamn, preposition, utländska ord och infinitivmärke representeras exakt samma beskrivningar. I kapitel 5 kommer dessa grupper och betydelsen av storleken på dem att jämföras och slutsatser dras om vilken betydelse de har.

Typer av morfologisk information som representeras i en tagguppsättning kan påverka en taggares prestation. Vilka taggar man väljer för att tagga ett ord har också betydelse beroende på information som finns i ett lexikon. Det föreligger också ett subjektivt val av vilka ord som kan anses tillhöra en viss ordklass.

4.1.3.2 Problem i mappningen

Vid mappningsprocessen uppkommer två typer av problem: dels problem med koherens i översättningen mellan tagguppsättningar, dels skillnader i taggning, t.ex. subjektiva val av taggar, olika lexikon som utgångspunkt samt feltaggning.

Vid översättningen konstateras ofta att taggarna inte står i ett 1:1-förhållande till varandra. Till exempel om en tagg med generella värden på attributen ska översättas till taggar med mer specifika värden. Det innebär att ett val måste göras mellan två eller fler taggar. Ofta kan det röra sig om egenskaper såsom numerus, genus eller species som behöver specificeras. I de flesta fall kan sådana problem lösas med någon form av kongruenstest, där information som behövs för att hitta rätt tagg kan finnas i närkontexten. Problem uppstår när kontexten inte innehåller sådan information, då kan de aktuella fallen eventuellt lösas genom manuell korrigering. I tabell 18 ovan finns ett exempel som automatiskt kan lösas med kongruenstest och i tabell 17 anges ett exempel där manuell korrigering är nödvändig.

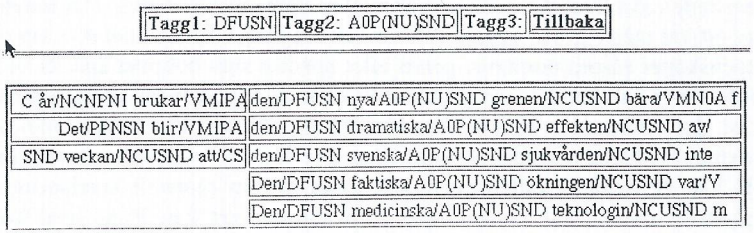
Trots att SUC genomgående har en tagguppsättning med mer specifik morfologisk information uppstår ändå 17 984 fall i mappningen av SUC, då SMDB använder mer specifika taggar för att beskriva ord än vad SUC gör. Detta är anmärkningsvärt just med tanke på att SUC:s tagguppsättning är större. Det omvända fallet hade varit mer troligt. Ett exempel på detta är "det sista nya Viggenplanet", där "nya" i SUC taggas med "`JJ POS UTR/`

NEU SIN DEF NOM", medan SMDB har två taggar som motsvarighet för ordet, nämligen "AQPUSND, AQPNSND".

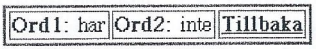
Val av olika taggar är problematiskt i mappningen därför att målet med översättningen till SMDB:s tagguppsättning innebär två saker, dels att alla textord ska ha taggar som är utformade med SMDB:s tagguppsättning, och dels att alla textord får den taggen, eller en av de möjliga taggar, som ordet har i SMDB.

Förutom bristande samstämmighet mellan val av taggar som har gjorts i de olika materialen SUC och SMDB, så är också felaktigheter i det taggade träningsmaterialet ett annat problem. Inkorrekt taggning får samma konsekvenser som det ovan beskrivna problemet med val av olika taggar. Feltaggningar av homografer beror sannolikt på att vid en manuell korrigering av automatiskt taggat material missas vissa felaktigheter (fel i taggning i SUC se exempel i tabell 2–tabell 4). Inkonsekvens, det vill säga att ord i samma kontext som har taggats på två eller fler sätt, upptäcks ganska lätt när konkordansverktyg används för att studera en text.

I denna studie har ett verktyg utvecklats där sökning är möjlig dels på en eller flera kombinationer av taggar och dels på ett eller flera kombinationer av ord; se figur 7 och figur 8 nedan. Inkonsekvens är den vanligaste typen av fel i SUC-korpusen. För exempel på inkonsekvent taggning, se exempel 5–exempel 6.



Figur 7. Gränssnitt till konkordansverktyg utvecklat inom avhandlingen för EPOS 1.0.



Uttalandet/NCNSND	har/VMIPA inte/RN0 fått/VMU0A några/DI(N
USNI Ulf-G/NP Gerdtham/NP	har/VMIPA inte/RN0 studerat/VMU0A orsake
NU)PN(ID) länderna/NCNPND	har/VMIPA inte/RN0 haft/VMU0A några/DI(N
Han/PP0SN	har/VMIPA inte/RN0 råd/NCNSNI att/CI läg
RT0 ,/FI men/CC han/PP0SN	har/VMIPA inte/RN0 haft/VMU0A tid/NCUSNI

Figur 8. Gränssnitt till konkordansverktyg utvecklat inom avhandlingen för EPOS 1.0.

I tabell 28 finns exempel på de mest frekventa typerna av skillnader (a)–(e) mellan SUC och SMDB. I (a) ges exempel på de tio vanligaste skillnaderna, där "andra" toppar listan med sammanlagt över 1500 förekomster. Skillnaden mellan tagguppsättningarna gäller då att SUC använder en adjektivtagg medan SMDB taggar ordet med någon av taggarna "MO*SND, V0N0A, V0M0A, PI*P*, DI*P, PI*S*, DI*S".

I (b) visas några exempel på typiska disambigueringsfel av homografer, gällande genus och kasus.

I grupp (c) finns ord som är typiska i talspråk, men ej skriftspråk och därför har dessa uteslutits i SMDB. Andra fel förekommer som ex. "blitt", vilket är felaktigt stavat. "francs", "dollars", är inte svenska ord, men har ändå taggats som nomen med pluralform. I korrekt svenska anges pluralform av dessa som "franc", "dollar". Om utländska former används bör dessa ord taggas som utländska ord.

I grupp (d) anges ord som tolkas på andra sätt i SUC och som därför ges andra taggar. Främst i denna grupp är nominaliserade adjektiv, exempelvis "medverkande", vilka alla är presens particip och har en ändelse på "-ande" eller "-ende". Dessa anges i SUC som nomen, medan de i SMDB är adjektiv.

Två ord som tolkas annorlunda är t.ex. "vänster" och "faktum". "Vänster" är taggat som adjektiv obestämd form i SUC, men har fått värdet "0" i SMDB, vilket innebär att det inte anses vara möjligt att urskilja bestämd och obestämd form. "Faktum" är ytterligare ett exempel som skiljer sig åt i species. I SUC taggas det med bestämd form, trots att det heter "faktumet" i bestämd form. Det är svårt att avgöra om det är ett taggningsfel. På liknande sätt taggas "centrum" med bestämd form och "preteritum", vissa ord med ändelsen "-um".

Vissa ord tolkas som adverb i SUC, trots att det kan vara svårt att förstå hur de som reguljära adverb skulle kunna vara en bestämning till ett verb eller ett adjektiv. Adverbialpartiklar finns inte i SMDB och därför taggas de som adverb. Vissa ord som "slut" och "reda" tolkas som partiklar i SUC, men de är inte adverb i SMDB och har därför inte tolkats som adverb utan nomen: "slut" (NCNSNI) och "reda" (NCUSNI).

Adverb kan i vissa flerordsenheter, t.ex. "i går", taggas så att "går" får en adverbtagg, men endast i de sammanhang då det gäller adverbet "i går". I SMDB taggas detta uttryck som preposition plus "NC0000" (oböjligt substantiv). Ett mål är därför att flerordsenheter som dessa i EPOS ska tolkas som sammanhängande kedjor och få endast en tagg, liknande det som har gjorts för proprier, för att undvika taggningsfel. Ett sådant arbete med flerordsenheter kan utföras som en fortsättning på denna studie.

Vissa skillnader finns i hur verbets imperativform och interjektioner används. I tabell 28 anges tre exempel där SUC taggar ord som interjektioner men SMDB har nomen, verb och adverb som möjliga alternativ. Ett annat exempel är "marsch", med motsatt förhållande, dvs. där det i SUC anges verb och imperativform, medan det i SMDB anges nomen eller interjektion.

Bokstäver tolkas också som förkortningar i SUC medan de i SMDB är nomen. Andra förkortningar är t.ex. "min" och "mån", vilka inte finns med i SMDB.

I grupp (e) finns några exempel som sannolikt är taggningsfel. Det börjar med några tveksamma adverb, vilka liksom de i grupp (d) är svåra att använda i syfte att bestämma verb eller adjektiv. Vidare finns ett par ord som får verbtaggar med infinitivform i stället för preteritumform. Ytterligare ett exempel är "skrivfel" vilket taggas som adjektiv i stället för nomen.

F.	Ord	SUC	SMDB	Mappad
(a) De tio vanligaste problemfallen:				
1199	andra	JJ POS UTR/NEU PLU IND/DEF NOM	MO*SND,V0N0A,V0M0A,PI*P*, DI*P,PI*S*,DI*S	AQP*PN*
1074	alla	DT UTR/NEU PLU IND/ DEF	NCUSNI,PI*P*,D0*P	DF*P
310	andra	JJ POS UTR/NEU SIN DEF NOM	MO*SND,V0N0A,V0M0A,PI*P*, DI*P,PI*S*,DI*S	AQP*SND
284	sista	JJ SUV UTR/NEU SIN/ PLU DEF NOM	NCUSNI,AQPN-SND,AQPUSND,AQP*PN*,AQP NSNI,AQPUSNI	AQS**ND

Tabell 28. Exempel på ord som skiljer sig i taggning mellan SUC och SMDB.

F.	Ord	SUC	SMDB	Mappad
240	båda	JJ POS UTR/NEU PLU IND/DEF NOM	NCUSNI,V0N0A,V0M0A,PF*P*, DF*P	AQP*PN*
139	något	AB	PINS*,DINS	R0
131	ett	PN NEU SIN IND SUB/ OBJ	MCNSNI,DINS	PINS*
111	anställda	NN UTR PLU IND NOM	AF0USND,AF0NSND,AF0*PN*, AQPUSND,AQPNSND,AQP*PN*	NCUPNI
97	Alla	DT UTR/NEU PLU IND/ DEF	NCUSNI,PI*P*,D0*P	DF*P
92	inre	JJ KOM UTR/NEU SIN/ PLU IND/DEF NOM	AQPUSNI,AQPN-SNI,AQPUSND,AQPNSND,AQP* PN*,R0	AQC**N*
(b) Fel i genus, kasus				
1	presi-dentval	NN UTR SIN IND NOM	NCNSNI,NCNPNI - genusfel	NCUSNI
1	misstagen	NN UTR PLU DEF NOM	NCNPND,AF0USNI - genusfel	NCUPND
1	mikrokos-mos	NN UTR SIN IND NOM	NCNSNI,NCNSGI - genusfel	NCUSNI
1	meters	NN UTR PLU IND NOM	NCUSGI,NCUPGI - kasusfel	NCUPNI
(c) Talspråk, eller felaktiga ordformer				
1	kalla	VB PRT AKT	AQPUSND,AQN-SND,AQP*PN*,V0N0A,V0M0A, NCUSNI	V0ISA
1	ja	PN UTR SIN DEF SUB	I,NCNPNI,NCNSNI	PFUSS
1	fråga	VB PRT AKT	NCUSNI,V0N0A,V0M0A	V0ISA
1	flytta	VB PRT AKT	V0N0A,V0M0A	V0ISA
1	fiska	VB PRT AKT	V0N0A,V0M0A	V0ISA
2	å	IE	NCNSNI,NCNPNI,NCUSNI,S,I	Z
1	blitt	VB SUP AKT	AQPNSNI,RP	V0U0A
1	francs	NN UTR PLU IND NOM	NCUSGI,NCUPGI	NCUPNI
1	dollars	NN UTR PLU IND NOM	NCUSGI,NCUPGI	NCUPNI
(d) Annan tolkning av ord ger andra ordklasser				
9	veder-börande	NN UTR SIN IND NOM	AQPUSNI,AQPNSNI, AQPUSND,AQPNSND, AQP*PN*,R0	NCUSNI
6	boende	NN UTR PLU IND NOM	AQPUSNI,AQPNSNI, AQPUSND,AQPNSND, AQP*PN*,R0,AP0**N*,NCNSNI	NCUPNI
5	Medver-kande	NN UTR PLU IND NOM	AP0**N*,R0,NCNSNI	NCUPNI
1	innersta	NN NEU SIN DEF NOM	AQS*PN*,AQSNSND,AQSUSND	NCNSND
1	cancers-juka	NN UTR SIN IND NOM	AQPUSND,AQPNSND,AQP*PN*	NCUSNI
33	vänster	JJ POS UTR SIN IND NOM	NCUSNI,AQP*SN0	AQPUSNI
36	faktum	NN NEU SIN DEF NOM	NCNSNI,NCNPNI	NCNSND

Tabell 28. Exempel på ord som skiljer sig i taggning mellan SUC och SMDB.

F.	Ord	SUC	SMDB	Mappad
1	genomar-betat	AB	V0U0A,AF0NSNI	R0
1	Garanterat	AB	V0U0A,AF0NSNI	R0
1	förvissat	AB	V0U0A,AF0NSNI	R0
44	slut	PL	NCNSNI,NCNPNI,NCUSNI, AQP0000,V0M0A	R0
66	reda	PL	NCUSNI,V0N0A,AQP0000, V0M0A	R0
53	går	AB	V0IPA,NC0000	R0
1	Herrans	JJ POS UTR/NEU SIN/PLU IND/DEF NOM	NCUSGD,NPNSGD	AQP**N*
1	marsch	VB IMP AKT	NCUSNI,I	V0M0A
1	Kryp	IN	NCNSNI,NCNPNI,V0M0A	I
1	Nu	IN	R0,NCNSNI	I
4	Gud	IN	NCUSNI,NPNSND	I
13	h	NN AN	NCNSNI,NCNPNI	Y
2	min	NN AN	NCUSNI,PSUS*	Y
(e) Felaktigt taggat				
1	cyklandes	AB	AP0**G*,NCNSGI	R0
2	väntat	AB	V0U0A,AF0NSNI	R0
3	bortsett	AB	V0U0A,AF0NSNI	R0
1	forsnings-centra	NN NEU SIN IND NOM	NCNPNI,NCNPND	NCNSNI
1	be	NN NEU SIN IND NOM	V0N0A,V0M0A	NCNSNI
3	fjärran	NN UTR SIN IND NOM	R0,AQPUSNI,AQPNSNI, AQPUSND,AQPNSND, AQP*PN*, NCNSNI	NCUSNI
4	litas	NN UTR SIN IND NOM	V0N0P,V0IPS	NCUSNI
3	kasus	NN UTR PLU IND NOM	NCUSNI,NCUSGI,NCN-SNI,NCNSGI,NCNPNI,NCNPGI	NCUPNI
1	hjälp	VB INF AKT	NCUSNI,V0M0A	V0N0A
2	var	VB INF AKT	NCNSNI,NCN-PNI,NCUSNI,AQP0000,R0,V0M0A,V0ISA,PIUS0,DIUS	V0N0A
1	misstagen	NN UTR PLU DEF NOM	NCNPND,AF0USNI	NCUPND
1	meters	NN UTR PLU IND NOM	NCUSGI,NCUPGI	NCUPNI
1	skrivfel	JJ POS UTR/NEU SIN/PLU IND/DEF NOM	NCNSNI,NCNPNI	AQP**N*

Tabell 28. Exempel på ord som skiljer sig i taggning mellan SUC och SMDB.

De mappningsproblem exemplifierats är en bråkdel av den faktiska mängden, ca 6400. Oavsett vilka problemen och felen i mappningen är, kan man vara säker på att det har stor betydelse för hur syntaktiska regler kommer att utformas vid träningen av en taggare. Ett exempel på detta är orden "våras" och "höstas" vilka i SUC taggas som "NC0000", medan de i SMDB taggas som "V0N0P, V0IPS, V0N0D, V0IPD, V0M0D, NCU000"

respektive "V0N0P, V0IPS, NCU000". Följden av att orden inte kan mappas blir att det inte bildas några syntaktiska trigramregler vid träningen. Ordet "våras" taggades vid disambiguering nästan konsekvent fel. Därför är det viktigt att mappningen kan fullföljas och alla problemfall lösas. Fel kan korrigeras. I övriga fall där man i SUC har valt t.ex. en adjektivtagg och i SMDB finns ett pronomen eller bestämd artikel att välja mellan, måste lösas manuellt. I avsnitt 7.1 redovisas ansatser till att förbättra mappningen av fallet med "höstas".

Frågan är varför det finns så många skillnader i SUC:s och SMDB:s tagguppsättning. Svaret är sannolikt att de har konstruerats för olika syften. SUC:s tagguppsättning har förmodligen anpassats för att så väl som möjligt beskriva alla token i en text och dessutom skapats utgående ifrån innehållet i lexikonet i SWETWOL. SMDB:s tagguppsättning har enbart skapats med utgångspunkt från SMDB och SAOL. Att den senare tagguppsättningen inte kan anpassas hur som helst för att täcka talspråk och dylikt i en korpus beror på de riktlinjer och normerande språk som dessa lexikaliska källor står för. Enligt författarens mening bör det också så förbli, men med kompletterande källor till SMDB, vilka t. ex. beskriver fler förkortningar och talspråk etc.

Extraherade statistiska regler, som tillämpats i studien, ger inte alltid förväntat resultat p.g.a. de fel som finns i den taggade korpusen. Detta bör beaktas när man studerar kvaliteten på taggningen i EPOS. Om det fanns ett helt korrekt taggat material skulle representationen i disambigueringsreglerna och precisionen öka i själva taggningen. Diskussioner som har förts angående mappning är av allmänt intresse. och de som har utfört liknande arbeten känner säkerligen igen problemen. Vissa problem som har uppstått i mappningen är av mer generell karaktär, och vissa andra av mer språkspecifik karaktär.

4.2 Disambigueringsmetod - EPOS-En Probabilistisk Ordklasstaggare för Svenska

I detta avsnitt redogörs för uppbyggnaden av det verktyg, EPOS, som har konstruerats i syfte att användas för fullständig taggning. Modellen EPOS med dess ingående moduler redogörs för i avsnitt 4.2.2, samt 4.2.2.3–4.2.2.7 för respektive moduler. I avsnitt 4.2.2.2 redovisas vilka tekniker inom ML som har använts. Metoden som har använts för disambiguering av homografer beskrivs i relation till den tidigare redovisningen av metoder (avsnitt 2.1.3.2). Slutligen ges exempel på problem som uppstod vid disambigueringen av homografer.

4.2.1 Bakgrund

Ordklasstaggning bör bedömas med avseende på sammanhang och syfte och därför är det viktigt att beskriva det sammanhang i vilket EPOS ingår.

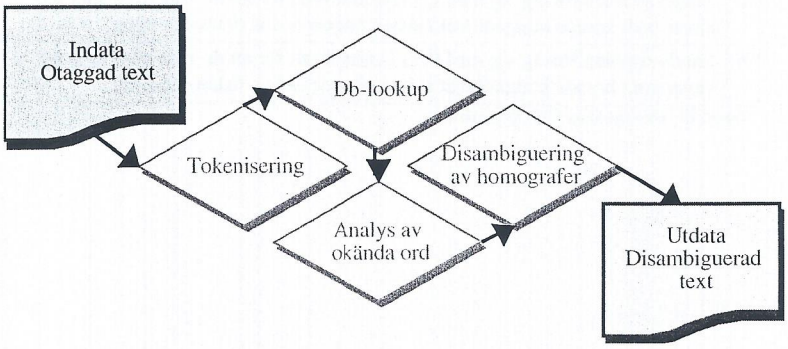
EPOS v. 2.0 skiljer sig från EPOS v. 1.0 eftersom v. 2.0 ingår i ett samarbete med målet att morfosyntaktiskt uppmärka texter i Språkbanken med utgångspunkt från SMDB. Därför modifierades innehåll i verktyget för att bättre följa den lexikalisk utformning av SMDB och utformning av resultat från disambiguering efter riktlinjer i uppmärkning av texter i Språkbanken. De resurser som finns i v. 1.0 är en tokeniserare, ett lexikon, en morfologisk suffixanalys och en disambigueringsmodul. I disambigueringsmodulen finns dessutom ett antal mindre moduler som undersöker vissa typer av ord, t.ex. adverbialpartiklar, adjektiv-nomen och andra fraser där koherens är viktigt. I EPOS v. 2.0 används endast modulen med morfologisk suffixanalys och disambigueringsmodulen. Anledningen till detta är att två av modulerna, Db-lookup och tokenisering, nu är modifierade och existerar som resurser i andra verksamheter. Tokeniseringen har ersatts med ett funktionsanrop till ett tokeniseringsprogram (Johansson Kokkinakis, 1996) som nu finns som ett av tillgängliga verktyg i Språkbanken. Db-lookup-modulen (annoteringsmodulen) bygger numera på SMDB, och har utvecklats av Cederholm (Berg et al. 2003), till skillnad från EPOS 1.0 där modulen byggde på SAOL11. Modulerna kommer redovisas kortfattat i det följande p.g.a. att de utgör moment i ordklasstaggning som påverkar de övriga modulerna i EPOS. Redovisningen omfattar även ersättningarna för de äldre modulerna.

4.2.2 EPOS-modellen

EPOS-modellen har använts för ordklasstaggning av texter i Språkbanken. Det står för "En probabilistisk ordklasstaggare för svenska". Dess design följer dagens normer för ordklasstaggningsverktyg (McEnery och Wilson, 1997). Dess innehåll och uppbyggnad beskrivs modulvis i följande avsnitt. Inledningsvis ges en generell beskrivning av innehåll och design och därefter följer modellens uppbyggnad.

4.2.2.1 Design och innehåll

McEnery och Wilson (1997) har angett nödvändiga komponenter för ett funktionellt ordklasstaggningssystem. EPOS skiljer sig inte nämnvärt från ett sådant system, och kan beskrivas som en modell med flera ingående moduler vilka alla har olika uppgifter i ordklasstaggning av en text. Modellen och de ingående modulerna illustreras i figur 9.



Figur 9. EPOS-modellen.

Vid en genomgång av alla moduler i modellen kan taggningsprocessen beskrivas stegvis med funktion och exempel för respektive steg.

Indata, exempel (a), är löpande, otaggad text vilken skickas vidare till första modulen som är *tokenisering*. Tokenisering kan beskrivas som en separering av ordsekvenser till enskilda tokens. Ett "token" motsvarar inte ett ortografiskt tecken utan ett eller flera ord eller ett interpunktionstecken, exempel (b). *Db-lookup* är den andra modulen vilken tar hand om den tokeniserade texten och för varje ord slår upp alla ord och deras möjliga taggar i den lexikaliska databasen, SMDB, exempel (c). Homografer anges i exemplet radseparerade fast med identiska id-nummer. Texten skickas sedan vidare till den tredje modulen vilken utför en *morfologisk analys* av förekommande okända ord, exempel (d). Modul fyra består av en

disambigueringsfunktion, vilken till 99% kan disambiguera homografa ord i texten. Korrektheten ligger på ca 94,5-96,5%, exempel (e). Detta är den slutliga utdatan från systemet.

I tabell 29 visas ett exempel på stegen i textbearbetningen och dess slutresultat. Exempel (a) och (b) skiljer sig endast genom att det senare fallet är tokeniserat. I (c) visas resultat från en "annoterings(lookup)-funktion" (Berg et al. 2003). Information som anges är 'tnr', tokennummer, 'atype', analystyp eller lookupinformation, 'msd', SMDB-taggar, 'lemma', lemmainformation och 'orth', ortografisk information. De fall som innehåller okända ord analyseras vidare i exempel (d). Slutligen efter disambiguering blir resultatet som i exempel (e) och (f). Det sista exemplet är en förenklad version av (e) som har tillkommit för att förbättra läsbarheten.

	Exempel
a	Saudi-Arabien försåg sig med CSS2-missilerna för att de ville göra känt att envar som hotade kungadömet måste göra en seriös riskbedömning.
b	Saudi-Arabien försåg sig med CSS2-missilerna för att de ville göra känt att envar som hotade kungadömet måste göra en seriös riskbedömning .

Tabell 29. Exempeldata i EPOS-modellen.

	Exempel
c	<ana tnr='836' atype='0' orth='Saudi-Arabien'/>
	<ana tnr='837' atype='1' msd='V0ISA' lemma='förse.1' orth='försåg'/>
	<ana tnr='838' atype='1' msd='PF00O' lemma='sig.1' orth='sig'/>
	<ana tnr='839' atype='1' msd='NCUSNI' lemma='med.2' orth='med'/>
	<ana tnr='839' atype='1' msd='R0' lemma='med.1' orth='med'/>
	<ana tnr='839' atype='1' msd='S' lemma='med.1' orth='med'/>
	<ana tnr='840' atype='0' orth='CSS2-missilerna'/>
	<ana tnr='841' atype='1' msd='AQPUSNI' lemma='för.5' orth='för'/>
	<ana tnr='841' atype='1' msd='CC' lemma='för.3' orth='för'/>
	<ana tnr='841' atype='1' msd='NCUSNI' lemma='för.4' orth='för'/>
	<ana tnr='841' atype='1' msd='R0' lemma='för.2' orth='för'/>
	<ana tnr='841' atype='1' msd='S' lemma='för.1' orth='för'/>
	<ana tnr='841' atype='1' msd='V0IPA' lemma='föra.1' orth='för'/>
	<ana tnr='841' atype='1' msd='V0M0A' lemma='föra.1' orth='för'/>
	<ana tnr='842' atype='1' msd='CS' lemma='att.1' orth='att'/>
	<ana tnr='842' atype='1' msd='Z' lemma='att.2' orth='att'/>
	<ana tnr='843' atype='1' msd='DF*P' lemma='den.2' orth='de'/>
	<ana tnr='843' atype='1' msd='PF*PS' lemma='den.1' orth='de'/>
	<ana tnr='844' atype='1' msd='V0ISA' lemma='vilja.2' orth='ville'/>
	<ana tnr='845' atype='1' msd='NCNSNI' lemma='göra.1' orth='göra'/>
	<ana tnr='845' atype='1' msd='V0N0A' lemma='göra.2' orth='göra'/>
	<ana tnr='846' atype='1' msd='AQPNSNI' lemma='känd.1' orth='känt'/>
	<ana tnr='846' atype='1' msd='RP' lemma='känd.1' orth='känt'/>
	<ana tnr='846' atype='1' msd='V0U0A' lemma='känna.1' orth='känt'/>
	<ana tnr='847' atype='1' msd='CS' lemma='att.1' orth='att'/>
	<ana tnr='847' atype='1' msd='Z' lemma='att.2' orth='att'/>
	<ana tnr='848' atype='1' msd='PIUS*' lemma='envar.1' orth='envar'/>
	<ana tnr='849' atype='1' msd='CC' lemma='som.1' orth='som'/>
	<ana tnr='849' atype='1' msd='CS' lemma='som.1' orth='som'/>
	<ana tnr='849' atype='1' msd='PH**0' lemma='som.2' orth='som'/>
	<ana tnr='849' atype='1' msd='R0' lemma='som.3' orth='som'/>
	<ana tnr='850' atype='1' msd='AF0*PN*' lemma='hota.1' orth='hotade'/>
	<ana tnr='850' atype='1' msd='AF0NSND' lemma='hota.1' orth='hotade'/>
	<ana tnr='850' atype='1' msd='AF0USND' lemma='hota.1' orth='hotade'/>
	<ana tnr='850' atype='1' msd='V0ISA' lemma='hota.1' orth='hotade'/>
	<ana tnr='851' atype='1' msd='NCNSND' lemma='kungadöme.1' orth='kungadömet'/>
	<ana tnr='852' atype='1' msd='NCNPNI' lemma='måste.2' orth='måste'/>
	<ana tnr='852' atype='1' msd='NCNSNI' lemma='måste.2' orth='måste'/>
	<ana tnr='852' atype='1' msd='V0IPA' lemma='måste.1' orth='måste'/>
	<ana tnr='852' atype='1' msd='V0ISA' lemma='måste.1' orth='måste'/>

Tabell 29. Exempeldata i EPOS-modellen.

	Exempel
c	<ana tnr='853' atype='1' msd='NCNSNI' lemma='göra.1' orth='göra'/> <ana tnr='853' atype='1' msd='V0N0A' lemma='göra.2' orth='göra'/> <ana tnr='854' atype='1' msd='DIUS' lemma='en.2' orth='en'/> <ana tnr='854' atype='1' msd='MCUSNI' lemma='en.1' orth='en'/> <ana tnr='854' atype='1' msd='NCUSNI' lemma='en.5' orth='en'/> <ana tnr='854' atype='1' msd='PIUS*' lemma='en.3' orth='en'/> <ana tnr='854' atype='1' msd='R0' lemma='en.4' orth='en'/> <ana tnr='855' atype='1' msd='AQPUSGI' lemma='seriös.1' orth='seriös'/> <ana tnr='855' atype='1' msd='AQPUSNI' lemma='seriös.1' orth='seriös'/> <ana tnr='856' atype='1' msd='NCUSNI' lemma='riskbedömning.1' orth='riskbedömning'/>
d	Okänt ord: Saudi-Arabien Ändelse: "abien" Föreslagen tagg: NPNSND Sannolikhet: 4.0957756170286e-03 Okänt ord: CSS2-missilerna Ändelse: "lerna" Föreslagen tagg: NCUPND Sannolikhet: 3.37901488404859e-02

Tabell 29. Exempeldata i EPOS-modellen.

	Exempel
e	<ana tnr='836' msd='NPNSND' lemma='-.-' atype='0' dtype='9' orth='Saudi-Arabien'/> <ana tnr='837' msd='V0ISA' lemma='förse.1' atype='1' dtype='00' orth='försåg'/> <ana tnr='838' msd='PF00O' lemma='sig.1' atype='1' dtype='00' orth='sig'/> <ana tnr='839' msd='S' lemma='med.1' atype='1' dtype='2' orth='med'/> <ana tnr='840' msd='NCUPND' lemma='-.-' atype='0' dtype='9' orth='CSS2-missilerna'/> <ana tnr='841' msd='S' lemma='för.1' atype='1' dtype='2' orth='för'/> <ana tnr='842' msd='CS' lemma='att.1' atype='1' dtype='1' orth='att'/> <ana tnr='843' msd='PF*PS' lemma='den.1' atype='1' dtype='12' orth='de'/> <ana tnr='844' msd='V0ISA' lemma='vilja.2' atype='1' dtype='00' orth='ville'/> <ana tnr='845' msd='V0N0A' lemma='göra.2' atype='1' dtype='3' orth='göra'/> <ana tnr='846' msd='V0U0A' lemma='känna.1' atype='1' dtype='3' orth='känt'/> <ana tnr='847' msd='CS' lemma='att.1' atype='1' dtype='12' orth='att'/> <ana tnr='848' msd='PIUS*' lemma='envar.1' atype='1' dtype='00' orth='envar'/> <ana tnr='849' msd='PH**0' lemma='som.2' atype='1' dtype='13' orth='som'/> <ana tnr='850' msd='V0ISA' lemma='hota.1' atype='1' dtype='3' orth='hotade'/> <ana tnr='851' msd='NCNSND' lemma='kungadöme.1' atype='1' dtype='00' orth='kungadömet'/> <ana tnr='852' msd='V0IPA' lemma='måste.1' atype='1' dtype='2' orth='måste'/> <ana tnr='853' msd='V0N0A' lemma='göra.2' atype='1' dtype='12' orth='göra'/> <ana tnr='854' msd='DIUS' lemma='en.2' atype='1' dtype='12' orth='en'/> <ana tnr='855' msd='AQPUSNI' lemma='seriös.1' atype='1' dtype='12' orth='seriös'/> <ana tnr='856' msd='NCUSNI' lemma='riskbedömning.1' atype='1' dtype='10' orth='riskbedömning'/> <ana tnr='857' msd='FE' lemma='-.-' atype='10' dtype='10' orth='.'/>
f	Saudi-Arabien/NPNSND försåg/V0ISA sig/PF00O med/S CSS2-missilerna/NCUPND för/S att/CS de/PF*PS ville/V0ISA göra/V0N0A känt/V0U0A att/CS envar/PIUS* som/PH**0 hotade/V0ISA kungadömet/NCNSND måste/V0IPA göra/V0N0A en/DIUS seriös/AQPUSNI riskbedömning/NCUSNI ./FE

Tabell 29. Exempeldata i EPOS-modellen.

En mer detaljerad beskrivning ges av de olika modulerna och komponenterna i EPOS. Tokenisering beskrivs i avsnitt 4.2.2.3, Annotering (Db-lookup) i avsnitt 4.2.2.4, morfologisk analys i avsnitt 4.2.2.5 och disambiguering av homografer i avsnitt 4.2.2.6, samt gränssnitt i avsnitt 4.2.2.7. Först beskrivs själva konstruktionen av verktyget och metoder för uppbyggnad i avsnitt 4.2.2.2.

4.2.2.2 Machine Learning vid konstruktion av EPOS

För att kunna utföra disambiguering av homografer samt morfologisk analys av okända ord är morfosyntaktiska regler för ordklasstaggar en förutsättning. Regler kan ha olika utseenden dels beroende på hur de används och dels beroende på hur de har skapats. I avsnitt 2.1.3.2 redogjordes både för disambigueringsmetoder, metodiskt och utseendemässigt, samt olika angreppssätt för att skapa regelmaterial.

Tillvägagångssättet vid disambiguering och morfologisk analys med EPOS har, innan den fick sin nuvarade form, bestått av handskrivna regler samt använt sig av Brills regelbaserade taggare. Försöket med handskrivna regler pågick under en tvåmånadersperiod och resulterade i ca 900 regler med 50% täckning på ca 1500 ord. Förutom låg prestation fanns nackdelen att detta angreppssätt kräver lingvistiska ställningstaganden som inte låg inom ramen för denna studie. Detta är skillnaden mellan ett lingvistiskt angreppssätt vid ordklasstaggning och ett datorstyrt angreppssätt (Voutilainen 1999). Däremot antogs att användning av Brill tagger skulle leda till en bättre förutsättning för disambigueringuppgiften. Taggarens prestanda visade sig också vara relativt låg, då den kunde uppmätas till ca 93% korrekt disambiguerad text. De flesta normalpresterande taggare idag klarar av ca 95%, beroende på hur specifik tagguppsättning som används. De ovan nämnda försöken resulterade i att EPOS idag består av en ny stokastisk disambigueringsmodul. Modulen har skapats med information automatiskt extraherad från morfosyntaktiskt uppmärkt korpusmaterial. Valet av de tekniker och metoder som har använts vid skapandet av taggaren beror dels på vilka resurser som har funnits tillgängliga under arbetets gång och dels på vad som tidigare hade gjorts inom området på institutionen och i övrigt vad gäller svenska språket.

För att få bästa möjliga underlag och öka förutsättningarna för framgångsrik ordklasstaggning och för att få en jämn fördelning mellan både lexikalisk och kontextuell information representerad i en korpus har probabilistiska värden från båda sidorna återskapats i automatiskt extraherade regler. De automatiska metoder som har använts i konstruktionen av EPOS är tekniker inom machine learning. Syftet är att automatiskt framställa

regelkomponenter. Resultatet bestående av olika typer av probabilistiska regler, skapades genom extrahering från SUC-materialet. Inlärningen kan klassificeras dels som *Case-Based Learning* därför att både tekniker som beskriver likhet eller närmaste motsvarighet till ett sökt regelmönster tillämpas Den kan också klassificeras som *regelinduktion* eftersom tekniker som generaliserar när sökt regelmönster saknas tillämpas.

Hur väl EPOS stämmer överens med ett inducerat inlärningssystem liknande det Daelemans strukturerar, i avsnitt 2.1.3, anges stegvis nedan;

Övervakning behövs i begränsad omfattning. Endast specificering är nödvändig beroende på vilken information som ska insamlas, t.ex. trigram för disambiguering eller morfologisk information för analys av okända ord.

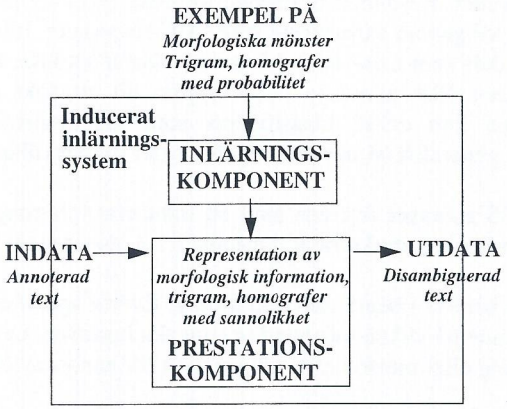
Representation av *indata* består av en tupel bestående av ett ord, en tagg, plus statistisk information om sannolikhet. Dessutom ingår ofta någon form av kontext, antingen som omgivande ord eller omgivande taggar.

Utdata består av en eller flera taggar.

Inlärningen sker genom att ny information läggs till gammal information. Inlärningstekniken är lat, dvs. varje exempel studeras men endast ny information läggs till.

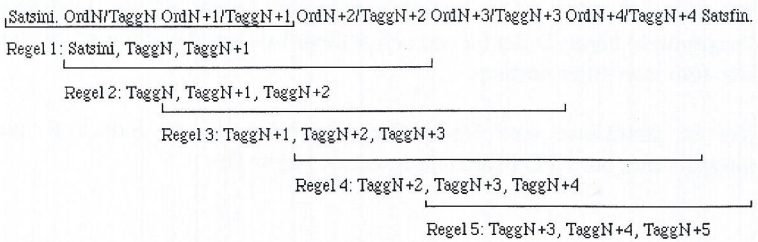
Toleransnivån för fel i inlärningsexempel ligger på en medelnivå. Programmet hanterar det mesta, och stannar inte om det stöter på ett mönster som inte följer normen.

Vid en jämförelse med Daelemans illustration av ett inducerat inlärningssystem, har EPOS följande utseende (figur 10).



Figur 10. Beskrivning av ett inducerat inlärningssystem i EPOS, Daelemans (1999).

Praktiskt kan extraheringen av kontextuell information beskrivas på följande sätt. Trigram konstrueras genom uppdelning av en morfosyntaktiskt taggad mening (se figur 11). En regel, i trigramformat, består av tre komponenter, dvs. tre på varandra följande taggar, tillsammans med ett sannolikhetsmått. Satsen i figur 11 genererar 5 regler, där antalet är proportionellt till antal ord i satsen, dvs. 5. Dessutom tillkommer ytterligare två regler med information om satsinitiala och satsfinala ord eller taggar.



Figur 11. Uppdelning av en taggad mening i trigram i träningen av EPOS.

En fråga som uppstår vid konstruktionen av regler av den ovan nämnda typen är vilken slags morfosyntaktisk information som ska ingå. Brill (1994) har experimenterat med träningsdata i korpusar. Han har använt både korpusar med ambiguöst material och med disambiguerat material. Brill (1995) har dragit slutsatsen att en kombination mellan båda typer av korpusar ger bäst resultat. Vanligast är dock att taggarna i träningskorpusen är disambiguerade, vilket också är fallet i EPOS.

4.2.2.3 Tokenisering

Tokeniseringen är en modul som innehåller ett program "Tokenize" som tokeniserar svensk text. Tokeniseringen är nödvändig för i stort sett alla tänkbara bearbetningar som kan göras av naturligt språk. Om en text inte har uppdelats i teckensekvenser försvåras former av vidare bearbetnings- eller analysprogram då det måste avgöras vilka sekvenser som faktiskt är ord, förkortningar, flerordsenheter och vilka som är interpunktionstecken.

Uppdelning av tecken är ett problem i dagens bearbetning av texter. Det är oftast svårt för olika parter att komma överens om definitionen av vad ett token är. Tokenisering kan definieras på följande sätt:

> Tokenization means dividing up the input text, which to a computer is just one long string of characters, into subunits, called *tokens*. Token means the individual appearance of a word in a certain position in a text. (Grefenstette, 1999 i van Halteren 1999:117)

Problem som förekommer är t. ex. ambiguitet vid ett ord med efterföljande punkt. Ordet eller förkortningen kan förekomma både med och utan efterföljande punkt. Interpunktionstecknet kan markera slutet av en sats eller vara en förkortning. Det underlättar därför om det finns en lista med vanliga förkortningar vid tokenisering. Ett annat problem är flerordsenheter. Ska "i dag" tolkas som en eller två enheter. Alternativen är "i_dag" eller "i" och "dag".

I EPOS 1.0 användes speciella förkortningslistor och listor över flerordsenheter för analys av dessa, men i version 2.0 används endast den information som finns i SMDB, vilken innehåller information om 388 förkortningar. Flerordsenheter representerades i den tidigare versionen av EPOS som listor, och i version 2.0 anges vanligt förekommande, t.ex. "idag" och "i dag", i SMDB, (278 stycken).

SMDB innehåller också 534704 egennamn och 525 utländska ord, (Berg et al. 2003).

4.2.2.4 Taggning med hjälp av lookup-metod och fullformsdatabas

Den första taggningen, som görs i en "annoteringsmodul", går ut på att slå upp varje ord i en tokeniserad text i SMDB och returnera alla möjliga ordklasser för ordet i fråga. Om ordet inte finns i databasen markeras det som okänt. Det är i detta skede en fördel att vänta med morfologisk analys av okänt ord p.g.a. att de omgivande orden i kontexten fortfarande inte är disambiguerade vilket medför att de ännu inte har "säkra" ordklasstaggar. Information som finns i SMDB är bl.a. ordens grundform kopplade till dess fullt utvecklade böjningsformer; se exempel i tabell 30 nedan.

Lemma (grundform+idnummer)	Böjningsformer	Avstavning	Grupp	Tagg
tiger.1	tiger	tiger	16	NCUSNI
	tigers	tigers	16	NCUSGI
	tigern	tigern	16	NCUSND
	tigerns	tigerns	16	NCUSGD
	tigrar	tigr+ar	16	NCUPNI
	tigrars	tigr+ars	16	NCUPGI
	tigrarna	tigr+ar+na	16	NCUPNI
	tigrarnas	tigr+ar+nas	16	NCUPGD

Tabell 30. Ett lemma och dess böjningsformer, från SMDB.

För lemmat "tiger.1" anges alla böjningsformer, information om avstavning med lämpliga brytningspunkter, böjningsgrupp samt tagg.

Böjningsformerna är automatiskt skapade med utgångspunkt från en uppsättning böjningskategorier, utvecklade av Sture Berg och Kaisa Samuelson, se Berg et al. (1988) om böjningskategorier samt tidigare avsnitt om SMDB (avsnitt 3.1.1). För mer information angående skapandet av SMDB, se Berg et al. (2003).

I EPOS 1.0 finns en motsvarande lexikalisk resurs, byggd på version 11 av SAOL samt de ovan nämnda böjningskategorierna, utvecklad inom Språkbanken, se Kokkinakis et. al. (1996). Den tidigare versionen innehåller ca 750.000 lemmatiska fullformer medan SMDB innehåller ca 1,2 miljoner fullformer.

Syftet med ett lexikon och hur det ska interagera med indata som tokeniserats anger Schiller och Karttunen (1999) på följande sätt:

> "The purpose of a tagger lexicon is to provide for each token of a text its potential properties. [...] At the input side of the lexicon component, the tokens which correspond to lexicon entries must be compatible with the tokenizer output." (Shiller och Karttunen, 1999 i van Halteren 1999:135).

Vid övergången mellan utdata från tokeniseringsmodulen och indata till Annoteringsmodulen måste ett textord ha samma form; det gäller främst flerordsenheter men även förkortningar samt ord vilka är sammansatta med bindestreck, t.ex. "Göteborg-Stockholm".

Vidare påpekar Schiller och Karttunen (1999) att innehållet i lexikonet beror på metod och tillvägagångssätt hos den taggare som ska användas:

> "The tagger lexicon may also include additional information about the inflection and the lemma or lemmas that word is an instance of. Other information may be necessary because of the tagger method, e.g. frequency information. [...] The structure or encoding of the lexicon depend of course on the individual tagger methods and programs." (Shiller och Karttunen, 1999 i van Halteren 1999:135).

Annan information i lexikonet än den om grundform samt böjningsmönster, som t.ex. frekvensinformation el. liknande, ges i den morfologiska modulen och disambigueringsmodulen eftersom det är i de modulerna som den används.

4.2.2.5 Morfologisk analys av okända ord

Den tredje modulen i taggningsverktyget är den som utför en morfologisk analys på okända ord. Denna modul arbetar dels med kontextuella regler och dels med det morfologiska utseendet av ett ord för att hitta en möjlig ordklass.

Information som används för analysen har skapats genom automatisk extrahering av dels kontextregler med suffixinformation, dels enbart suffixinformation. I båda fallen anges också sannolikhetsinformation. Anledningen till detta val av metod är att den är tidsbesparande samt mycket effektiv. Autentiska exempel på morfologisk ändelsevariation ger ett stort utbud av användbara analysregler. För att manuellt åstadkomma motsvarande material genom regelformation av vanliga konstruerade morfologiska suffix i kontext skulle oerhörda tidsresurser krävas. Framgången med denna automatiska metod visar sig t.ex. som god kvalitet vid analys av utländska ord, förkortningar och facktermer.

Modulen arbetar med ett antal tester och beroende på om det finns regler som stämmer överens med givet suffix och kontext rapporteras en gissning på ordklasstillhörighet. I annat fall görs nya tester med mer begränsad kontextuell information. Tekniken som används vid analys är inte morfologisk i den betydelsen att man separerar ett ord i stam och ändelse; utan snarare undersöks ordets 'tekniska suffix', som definieras som de *n* sista tecknen (och alltså inte är en äkta morfologisk ändelse). Det är lättare att hitta en lämplig motsvarighet bland automatiskt genererade tekniska suffix, om en del av stammen på lemmat dessutom finns med. Resultatet av ett antal tester med olika antal tecken i suffixet visar att just 5 tecken ger träff med störst säkerhet. Formeln som används för analys av suffix i EPOS, dels med och dels utan kontext, anges i Ekv (7) och Ekv (8) nedan.

$$P\langle T_i|W_{ei},(T_{i-1},T_{i+1})\rangle = \frac{|W_{ei},T_i,(T_{i-1},T_{i+1})|}{|W_{ei},(T_{i-1},T_{i+1})|}$$

Sannolikhetsformel för okända ord, med suffix och kontext i EPOS. Ekv (7)

Formeln i Ekv (7) utläses så, att sannolikheten P för att ett ord W_i som har ändelsen e får taggen T_i, är summan av alla förekomster av ordet W_i med ändelsen e och taggen T_i med taggarna T_{i-1} och T_{i+1} i kontexten, dividerat med summan av alla förekomster av ordet W_i med ändelsen e och taggarna T_{i-1} och T_{i+1} i kontexten. Kontextinformationen anges relativt det aktuella ordet i.

$$P\langle T_i|W_{ei}\rangle = \frac{|W_{ei},T_i|}{|W_{ei}|}$$

Sannolikhetsformel för okända ord, med suffix, utan kontext i EPOS. Ekv (8)

Formeln i Ekv (8) utläses så att sannolikheten P för att ett ord W_i som har ändelsen e får taggen T_i är summan av alla förekomster av ordet W_i med ändelsen e och taggen T_i, dividerat med summan av alla förekomster av ordet W_i med ändelsen e.

Den form av sannolikhetsmått som har använts för den morfologiska analysen är relativ frekvens. Dessa är relativt enkla, men effektiva då de till stor del uppfyller sitt syfte. Andra möjliga formler som kan användas i detta sammanhang kommer att beskrivas, men prövning av dessa ingår inte i denna studie. En annan formel vanlig formel för att beräkna sannolikhet är Bayes teorem (BT), Ekv (9).

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} \qquad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

a) b)

Bayes teorem för beräkning av sannolikhet för en företeelse givet en annan, Manning & Schütze (1999:43). Ekv (9)

Formeln a) kan användas för att beräkna sannolikhet givet vissa värden; den andra formeln b) är dess motsvarighet uttryckt i termer som används inom mängdlära.

Den första formeln i Ekv (9) utläses så att sannolikheten för B givet A, är sannolikheten för A givet B multiplicerat med sannolikheten för B, dividerat med sannolikheten för A. Den andra formeln utläses så att sannolikheten för B givet A, är sannolikheten för snittet B och A, dividerat med sannolikheten för A; jfr. (Manning & Shütze, 1999).

Skillnaden mellan Bayes Teorem och relativ frekvens är att det i Bayes Teorem tas hänsyn till sannolikheten för B samt sannolikheten för A, förutom den sammanlagda sannolikheten för de två. I avhandlingen används formler som beräknar sannolikheten för de enskilda värdena A och B, och sannolikheten för B givet A, separat.

Brill (1999), anger ett vanligt sätt att beräkna lexikalisk sannolikhet i en trigram-taggare enligt Markov modellen Ekv (10):

$$P(word|tag) = \frac{count\ (word\ and\ tag)}{count\ (tag)}$$

Formel för lexikalisk sannolikhet enligt beräkning med "maximum likelihood". (Brill 1999, i van Halteren 1999:210). Ekv (10)

Enligt Brill kan Ekv (10) inte kan användas för att beräkna sannolikhet för okända ord eftersom det inte finns okända ord i en träningskorpus. Därför hävdar Brill att det är mer fördelaktigt att använda sig av ord och ordkombinationer i en träningskorpus som endast förekommer en gång, "one-count word", och utföra en maximum-likelihoodberäkning för dessa samt sedan använda sig av resultatet för att beräkna sannolikhet för okända ord. Formeln skulle då se ut som Ekv (11) nedan.

$$P(\text{one-count word}|tag)=\frac{count\ (\text{one-count word and tag})}{count\ (tag)}$$

Förslag till formel för lexikalisk sannolikhet för okända ord med "one-count"-frekvens, (Brill 1999, i van Halteren 1999:210). Ekv (11)

Brill anger dessutom ett annat förslag till formel för att beräkna sannolikhet för okända ord, enligt Weischedel et al. (1993), Ekv (12):

$$P(wi|t_j)= P(\text{one-count word}|t_j) \times P(\text{Capital Feature}|t_j) \times P(\text{Hyphenation}|t_j) \times P(\text{Endings}|t_j)$$

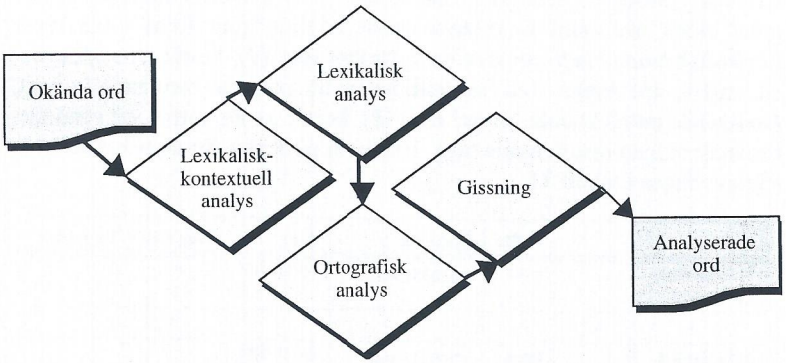
Uträkning av lexikalisk sannolikhet för okända ord, (Weischedel 1993, i van Halteren 1999:210). Ekv (12)

Brill (1999) förklarar formeln på följande sätt:

> ...where *Capital Feature* is a binary feature which is one if the first letter of the word is a capital letter and zero otherwise, the *Hyphenation* feature is a binary feature indicating whether the word has a hyphen in it and *Endings* is a j+1-valued feature indicating which (if any) of j suffixes a word has. (Brill 1999, i van Halteren 1999:210)

Att notera i Weischedels formel är att värden på de fyra olika informationstyperna kan bli '0'. Det innebär inte nödvändigtvis att hela ekvationen blir '0'. Då inga detaljer framgår av faktisk implementering av formeln kan man anta att informationstyper som inte har något värde förmodligen inte tas med i beräkningen.

I likhet med Weischedels ekvation, ingår i EPOS morfologiska modul att hänsyn tas till "Capital Feature", "Hyphenation" samt "Endings". Till skillnad från ekvationen i Ekv (12), används inte data om ord som endast förekommer en gång, utan i stället används data om alla existerande ord i träningskorpusen, då författarens åsikt är att detta kan skapa en jämnare och bättre fördelning av suffixdata som kan användas i analys av okända ord. Suffixanalys skiljer sig mellan svenska och engelska genom att svenskan har rikare och mer varierade ändelser än vad som finns i engelskan. Arbetsgången för analys av okända ord illustreras i figur 12.



Figur 12. Metod för analys av okända ord i den morfologiska modulen i EPOS.

Ett exempel med två ord på vardera sidan i omgivande kontext ges i exempel 11 och exempel 13, för analys av det okända ordet "CSS2-missilerna";

```
<ana tnr='838' atype='1' msd='PF00O' lemma='sig.1' orth='sig'/>
<ana tnr='839' atype='1' msd='NCUSNI' lemma='med.2' orth='med'/>
<ana tnr='839' atype='1' msd='R0' lemma='med.1' orth='med'/>
<ana tnr='839' atype='1' msd='S' lemma='med.1' orth='med'/>
<ana tnr='840' atype='0' orth='CSS2-missilerna'/>
<ana tnr='841' atype='1' msd='AQPUSNI' lemma='för.5' orth='för'/>
<ana tnr='841' atype='1' msd='CC' lemma='för.3' orth='för'/>
<ana tnr='841' atype='1' msd='NCUSNI' lemma='för.4' orth='för'/>
<ana tnr='841' atype='1' msd='R0' lemma='för.2' orth='för'/>
<ana tnr='841' atype='1' msd='S' lemma='för.1' orth='för'/>
<ana tnr='841' atype='1' msd='V0IPA' lemma='föra.1' orth='för'/>
<ana tnr='841' atype='1' msd='V0M0A' lemma='föra.1' orth='för'/>
<ana tnr='842' atype='1' msd='CS' lemma='att.1' orth='att'/>
<ana tnr='842' atype='1' msd='Z' lemma='att.2' orth='att'/>
```

Exempel 12. SMDB-annoterad text med homografer, okänt ord: "CSS2-missilerna"

```
<ana tnr='838' msd='PF00O' lemma='sig.1' atype='1' dtype='00' orth='sig'/>
<ana tnr='839' msd='S' lemma='med.1' atype='1' dtype='2' orth='med'/>
<ana tnr='840' msd='NCUPND' lemma='-.-' atype='0' dtype='9' orth='CSS2-mis-
silerna'/>
<ana tnr='841' msd='S' lemma='för.1' atype='1' dtype='2' orth='för'/>
<ana tnr='842' msd='CS' lemma='att.1' atype='1' dtype='1' orth='att'/>
```

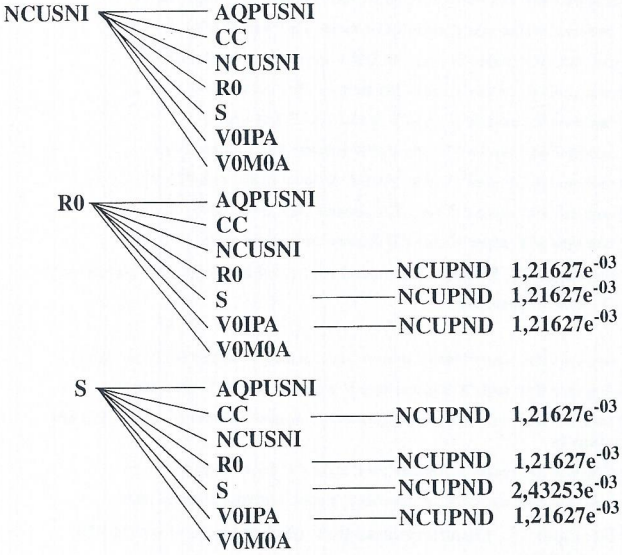
Exempel 13. Disambiguerad text, med analyserat okänt ord, med EPOS.

Lexikonuppmärkt text innehållande okända ord skickas till den morfologiska analysmodulen. Där görs först en analys av överensstämmelse mellan givet okänt ord samt frekvensbaserade trigram som finns i två typer. *Lexikalisk-kontextuell*, motsvarande formel Ekv (7), består av regler med ett suffix, tre taggar, ett sannolikhetsmått och en resulterande tagg. *Lexikalisk*, motsvarande formel Ekv (8), består av ett suffix, ett sannolikhetsmått och en resulterande tagg. Exempel på regler för de två typerna av trigram anges i tabell 31.

Typ	Suffix	Frekvens	Tagg	Tagg före	Tagg efter
1) Lexikalisk-kontextuell	lerna	$3.412352941e^{-03}$	NCUPND	S	S
2) Lexikalisk	lerna	$9.379014884e^{-03}$	NCUPND		

Tabell 31. Beskrivning av två modeller för trigram vid morfologisk analys.

För att uppnå högre precision i analysen tillämpas först typ 1 eftersom det ger en mer exakt träff med ingående kontext, och om inget resultat uppnås görs analys med typ 2. Frekvensmåtten skiljer sig mellan typerna. De är lägre på typ 1 eftersom det är en mer exakt variant, och högre på typ 2 eftersom det är en mer generell variant. För enkelhetens skull har endast en tagg före och efter i kontexten angivits. Vid homograf kontext som den ovan, blir det en alternativ lösning för varje möjlig kombination av de homografa kontexttaggarna, dvs. (NCUSNI-AQPUSNI, R0-AQPUSNI, S-AQPUSNI, NCUSNI-CC), etc. (Se figur 13 för en utveckling av möjliga kombinationer).



Figur 13. Antal möjliga kombinationer av taggar, med sannolikhetsvärden från EPOS.

Vid närmare jämförelse av de homografa kontexttaggarna framgår att sannolikhetsmåtten endast skiljer sig i ett fall, nämligen det med en prepositionstagg före och en prepositionstagg efter. Därför är detta den regel som är bäst lämpad för att ange lämplig tagg före ordet "CSS2-missilerna".

Om analys med typ 1 eller 2 inte hade gett något resultat tillämpas en ortografisk analys, då versaler och gemener på ordet undersöks samt dess kasusinformation. Om ordet t.ex. slutar på 's' är det antagligen genitivform av ett egennamn och annars om det inte slutar på 's', är det antagligen nominativformen av ett egennamn. Analysen försvåras avsevärt vid förekomst av ytterligare ett eller flera okända ord eller homografer i närkontexten. Det finns en gissningsdel i den morfologiska analysen vilken drar slutsatsen att ett ord är ett egennamn om ingen annan information existerar. Detta görs på grund av att sannolikheten för att nya okända ord är egennamn är betydligt större än utländskt ord, nomen eller adjektiv.

En utvärdering av analysmodulen för okända ord har gjorts i kapitel 5. Utvärderingen beskriver prestationen hos modulen med avseende på ordklasser. Exaktheten blir naturligtvis sämre när det inte bara finns en utan flera taggar att välja på i den närmaste kontexten. Recall på analys av okända ord är med EPOS 1.0 96%, samt EPOS 2.0 60,85%.

4.2.2.6 Disambiguering med hjälp av trigram, token och sannolikhet

Disambiguering av homografer är den mest krävande delen i ett taggningsverktyg. Homografseparering är nödvändig då det i genomsnitt är minst 50% av alla ord i en svensk text som kan tillhöra mer än en ordklass. I källor som t.ex. Nusvensk Frekvensordbok (1970) anges att homografin ligger högre, 64,5%, medan Karlsson (1992) anger värden på 57,3%. Homografin står i relation till hur ordklassbegreppet definieras. Ju fler homografer som finns representerade i ett lexikon desto fler homografer kan identifieras i en text. Därför är det naturligt att material från Språkdata såsom NFO och Press 65 innehåller högre homografimått eftersom de lexikon som används antagligen är de enda för svenska som innehåller så omfattande information om homografi, framför allt i SAOL.

Homograffrekvensen i den text som används i utvärderingen (kapitel 5) är ca 51%. Ett mått på den sammanlagda homograffrekvensen i den lexikaliska databasen är 17,1%. Denna frekvens fördelar sig på 1,24% externa homografer och 16,47% interna homografer. Frekvensen beräknas genom att man dividerar totala antalet taggar med antal ord. Man får då ett mått som beskriver hur många taggar ett ord i genomsnitt har. Dessa mått är starkt beroende av hur respektive lexikon och text är utformade.

Olika typer av texter har också varierande mått på homograffrekvens. En text som innehåller många ovanliga fackord och utländska ord bör innehålla mindre homografi än en text som innehåller många vardagliga ord och uttryck. SMDB används som lexikonresurs i taggningen och har förmodligen, i jämförelse med andra motsvarande lexikon, ett högre mått av homografi. Här bör noteras att det är svårare att disambiguera homografer med SMDB än med andra lexikon med färre och mindre homografa ord. Detta diskuteras vidare i utvärderingsavsnittet. Där föreslås ett nytt sätt att utvärdera disambiguering av texter med avseende på hur 'homograft' ett lexikon är och hur 'homograf' en viss text är.

Här följer en redovisning av val av metod och teknik vid skapandet av en disambigueringsmodul. Tillvägagångssättet var följande. Efter ett par ansatser med bl.a. handskrivna regler samt användande av Brill tagger, med olika, ej tillfredsställande, resultat, stod det klart att disambigueringsreglerna måste skapas på automatisk väg. Tekniker som används för att disambiguera homografer med hjälp av datatekniska metoder kan vara bl. a. regelbaserade eller stokastiska (avsnitt 2.1.3.2). Eftersom det inte fanns några givna preferenser vid val av metod eller teknik, testades en enkel matchningsteknik med sekvenser av taggar, s.k. *N-gram* som extraherats ur korpusmaterialet. Olika längd på sekvenserna gav resultat av olika kvalitet. Bäst resultat uppvisades vid längden tre, dvs. *trigram*, p.g.a. att

mönstren var tillräckligt långa för att hitta bra kontext för framgångsrik disambiguering och de var tillräckligt korta för att få flera olika alternativ vid regelmönstermatchning. Trigram är en relativt enkel representationsform av syntaktiska konstituenter jämfört med andra tekniker. De är dessutom mycket snabba att söka i om de lagras på ett utrymmesbesparande sätt med indexering i databasformat. CLAWS1 är ett exempel på ett system som använder N-grams för analys av engelska. Detta format på regler används även i HMM, vilket är en av de mest använda metoderna vid taggning idag.

I EPOS har stokastiska metoder används för morfologisk analys och disambiguering. Mer specifikt kan EPOS kallas en N-gram-taggare beroende på att den använder sig av regler bestående av ordklasskedjor eller -mönster. Dessa mönster kallas N-gram. Voutilainen (1999) ger exempel på tidiga användningar av N-grams, t.ex. CLAWS1, som var ett av de första taggningssystemen vilket till skillnad från sina föregångare inte använder sig av lingvistiska principer för att disambiguera homografer. CLAWS1 använde i stället statistisk korpusevidens för att bestämma korrekt tagg för ett ord. Taggningsverktyget har använt Brown-korpusen för att extrahera regler som bygger upp sannolikhetsmatriser, vilka innehåller information om likhet i kontext. CLAWS1 använder ordklassmönster med två taggar, sk. "bigram", till skillnad från tre taggar och "trigram", vilket används i EPOS. För att beräkna sannolikhet för ord, med hjälp av både lexikalisk och kontextuell sannolikhet påpekar Voutilainen följande om matriserna i CLAWS1;

> From a matrix of this kind, two kinds of probabilities can be calculated: lexical (the likelihood of word X representing category A) and contextual (the likelihood of category B occurring directly adjacent to category A) A probabilistic disambiguator should determine an optimal balance between lexical and contextual probabilities. The formula used in the final version of CLAWS1 is (Marshall 1987:42-65):

$$\frac{\textit{Frequency of the tag sequence A-B}}{\textit{Frequency of tag A x Frequency of tag B}}$$

Formel som anger balans mellan lexikalisk och kontextuell sannolikhet i CLAWS1, Marshall (1987:42-56). Ekv (13)

Vid utvärderingar har CLAWS1 uppvisat korrekthet på ca 96-97%. Andra kända N-gram taggare är "Part of Speech" av Church (1988) och "Volsunga" av DeRose (1988) med uppmätt korrekthet på 95-97% respektive 96%. Ytterligare användningar av bi- och trigram är Weischedel et al. för engelska (1993), Åström för svenska (1995) och Nagata för japanska

(1994).

Trigram består av en sekvens av tre ord och dess motsvarande taggar. Den vanligaste[15] uppställningen av taggarna är:

$$n-2, n-1, n$$

Mönster för trigram av ordklasser, (El-Beze et al. i van Halteren 1999:271-272). Ekv (14)

Ekv (14) ovan kan utläsas så, att analysen av en tagg i position *n* föregås av taggarna i positionerna *n-1* och *n-2*.

En annan uppställning av trigram används i EPOS, nämligen:

$$n-1, n, n+1$$

Mönster för trigram av ordklasser i EPOS. Ekv (15)

Mönstret i Ekv (15) skiljer sig från den tidigare angivna på det sättet att vid analys av ett ord n, anges både ett ord före, *n-1,* och ett ord efter, *n+1*. Denna omdisponering av trigram har framför allt gjorts i syfte att undvika sk. *backtracking*, vilket är frekvent förekommande i olika algoritmer, t.ex. vid HMM med Baum-Welch-algoritmen (Forward-Backward), och görs för att återskapa möjliga lösningar vid disambiguering som har blivit förbigångna p.g.a. låg sannolikhet etc. I den nya formen av trigram i EPOS hålls möjligheterna mer öppna genom att ett mönster delvis består av ett redan disambiguerat ord samt ett med alla möjliga taggar plus själva ordet som ska disambigueras och alla dess taggar med möjligheten att välja de taggar som är mest sannolika, se Ekv (16), där i,j varierar från 1 och uppåt.

$$(W_{n-1})\, t_{n-1},\ (W_n)\, t_i,\ (W_{n+1})\, t_j$$

Mönster för trigram med ord och ordklasser i EPOS. Ekv (16)

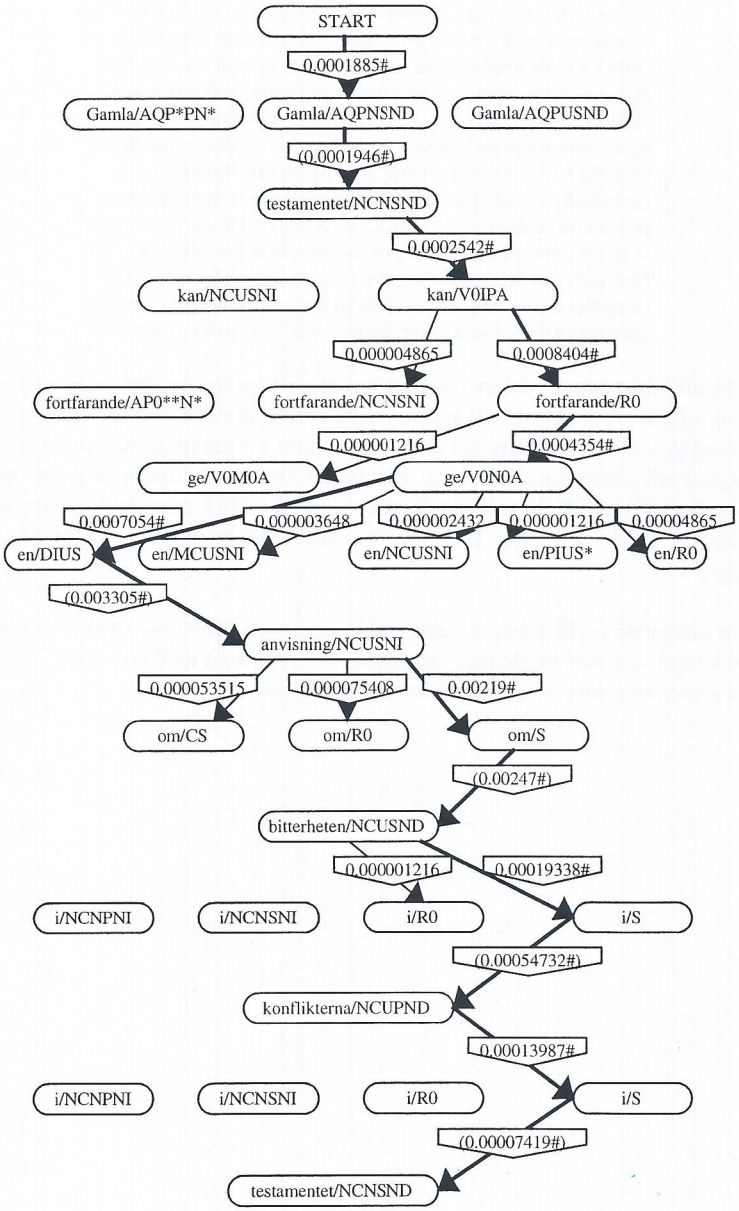
Viterbi-algoritmen (Viterbi 1967) använder sig av liknande tillvägagångssätt, nämligen att välja den optimala sökvägen, där beräkningen av mest sannolika taggar bildar en linjär funktion av utdatalängden. El-Beze (1999) förklarar algoritmen:

15. Det finns olika representationer i olika metoder som använder sig av trigram, vanligast kan antagligen (n-2, n-1, n) vara. Men med metoden HMM kan flera olika former av trigram förekomma, t. ex. (n-2, n-1, n), (n-1, n, n+1) och (n, n+1, n+2). Då används det trigram som ger bäst resultat i varje enskilt fall.

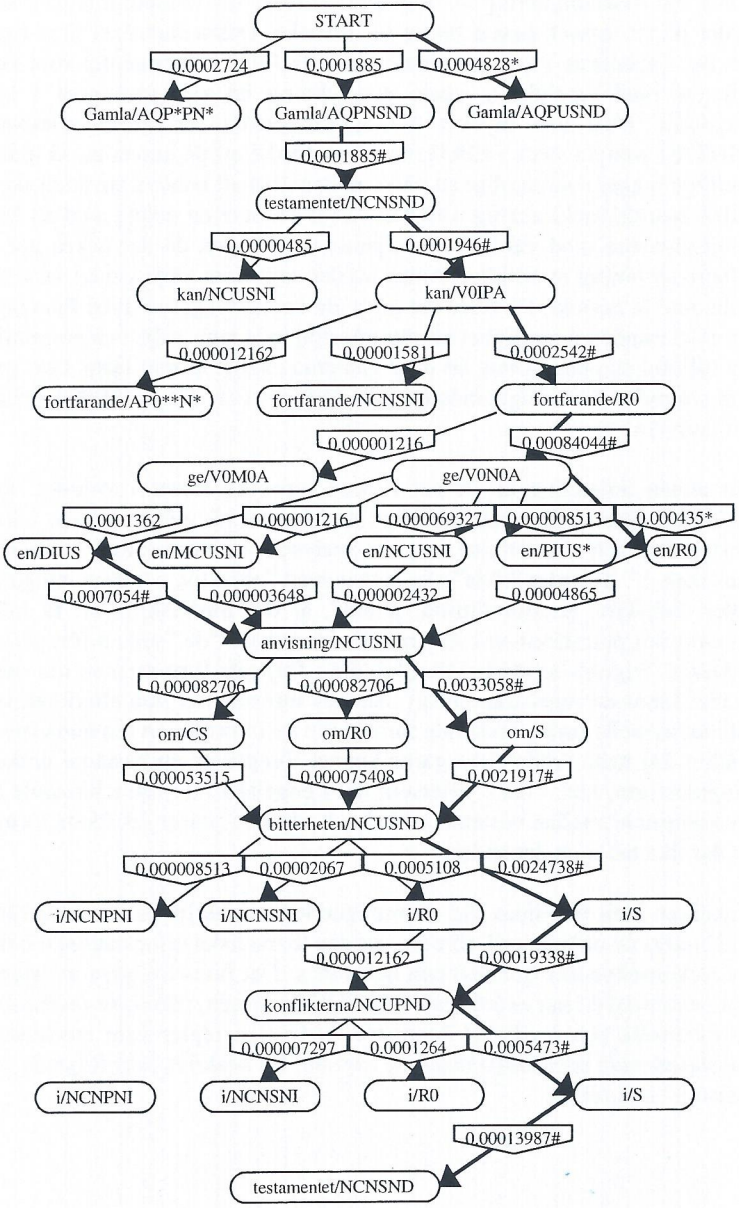
> It is based on the observation that, if you look at the best path that goes through a given state S at a given time t, the path is the concatenation of the best path that goes from the initial state to state S (while emitting symbols from 1 to t) with the best path that goes from state S to the final state (while emitting symbols from t+1 to the end). This is because the probability of a path going through state S is simply the product of the probabilities of these two parts (before and after S), so that the maximum probability of the global path is obtained when each part has a maximum probability. ... The number of computations involved is in the order of the number of transitions in the network... The best path that arrives at the final state of the model, while the complete output sequence has been emitted, will be the optimal path that we are looking for. (El-Beze et al. i van Halteren 1999:270)

För att illustrera fördelen med att använda trigram på det ovan angivna sättet jämfört med det mer traditionella sättet, görs en jämförelse av tillgängliga data av trigram, bestående av enbart kontextuell information, dvs taggar till disambiguerade ord. Resultatet blev att backtracking var nödvändig vid två tillfällen med trigram enligt Ekv (14), men det var inte nödvändigt med Ekv (15). Exempel på dessa analyser ges i figur 14 och figur 15.

En kommentar till bilderna: den valda analysen markeras med fet pil och den tagg som har högst frekvens med "#", felaktiga analyser med "*" och analyser som inte behöver utföras markeras med "()".



Figur 14. Exempel på analys med trigram enligt mönstret: n-1, n, n+1 i EPOS.



Figur 15. Exempel på analys med trigram enligt mönstret: n-2, n-1, n.

Förutom att den i EPOS använda formen av trigram ger uppenbara fördelar vid disambiguering av homografer, finns det dessutom tecken som tyder på att formen passar bättre för svenska. Dessa slutsatser dras p.g.a. att det i svenskan i t.ex. nominalfraser ofta finns komponenter som kongruerar med varandra i större utsträckning än vad som görs i t.ex. engelska. Med den mycket homograffrekventa lexikaliska databasen (SMDB) som används i EPOS, är det en fördel att då kunna ha så många möjliga taggar som möjligt att välja mellan utan att behöva använda sig av tidskrävande backtracking. Det kan därför ses som en poäng med att hålla möjligheterna med val av efterföljande tagg öppna då det också går att hindra spridning av följdfel. Antag att den tagg som väljs vid ett visst tillfälle inte är korrekt. Då finns det p.g.a. de öppna 'vägarna' med flera taggar, fortfarande en möjlighet att rätt efterföljande tagg väljs och sprindning av följdfel stoppas. Antag att möjligheterna istället är mer låsta, t.ex. genom två redan disambiguerade taggar i förkontexten. Då är följdfel svårare att undvika.

En annan anledning är att det i vissa typer av fraser i svenska, t.ex. nominalfraser verkar mer naturligt att både se på informationen i kontexten före och efter ett ord som ska analyseras. I ett exempel "... de röda hus som ..." då ordet "röda" ska taggas med "adjektiv, neutrum, singular" eller "adjektiv, neutrum/utrum, plural", är det fördelaktigt att få information om pluralformen i den bestämda artikeln "de" vilken föregår adjektivet. Trigram används inte lika ofta i följande format, men om det istället fanns en regel som enbart hämtade information från efterkontexten till det aktuella ordet, hade inte samma typ av information om numerus erhållits. Då hade "röda" lika gärna kunnat föregås av en bestämd artikel i singularform, t.ex. "det". Eftersom det i engelska inte finns liknande typ av kongruens mellan bestämd artikel och adjektiv passar det första formatet där lika bra som det andra.

Tekniken som tillämpas vid disambiguering är snarlik den som används vid analys av okända ord, förutom att den lexikaliska informationen består av teckensekvenser (tokens) och inte suffix. Det finns tre typer av trigram som automatiskt har extraherats från SUC-korpusen, nämligen *lexikaliska*, *kontextuella* och *lexikalisk-kontextuella*. Övriga regler som används, är *tokenbaserade* samt *suffixbaserade* trigram. De beskrivs i det följande med exempel och regler.

Typ	Ord	Sannolikhet	Tagg	Tagg/ord före	Tagg/ord efter
1) Lexikalisk	hus	0.00036488	NCNSNI	det	där
2) Lexikalisk-kontextuell	hus	0.00060813	NCNSNI	PSNS*	S
3) Kontextuell		0.130	NCNSNI	DINS	S
4) Tokenbaserad	hus	0.0090107	NCNSNI		
5) Suffixbaserad	*hus	0.0090107	NCNSNI		

Tabell 32. Exempel på tre regler med trigram i EPOS, plus två andra regler vilka används vid disambiguering.

Trigram av typ (3) är av mer generell än typ (1) och är därför mer frekventa och reglerna har följaktligen högre sannolikhet. Typ (2) kan klassificeras som något mitt emellan. Den lexikaliska typen (1) använder sig av enbart lexikalisk information, samt lexikalisk information i kontexten. Typ (2) använder lexikalisk information med kontextuell information om taggar. Typ (3) använder sig enbart av en möjlig tagg samt information om taggar i kontexten. Typ (4) bygger enbart på information om en lexikalisk enhet, t.ex. ett ord och typ (5) enbart på suffix, där asterisk (*) i sammanhanget betyder vilken teckensekvens som helst följt av "hus". Alla typer av regler anges med sannolikhet vilken bygger på relativ frekvens, se Ekv (17).

$$P(E_i) = \frac{|E_i|}{|E_{1-n}|}$$

Sannolikhet med relativ frekvens, Manning och Schütze (1999:197). Ekv (17)

I Ekv (17) anges sannolikheten för att händelsen E_i ska inträffa, som summan av alla förekomster av händelsen E_i dividerat med summan av alla händelser från 1 till n. Observera notationssättet, där 1-n i täljaren avser att beskriva alla händelser om det finns maximalt n händelser. Vid beskrivning av liknande formler i detta avhandlingsprojekt, där det i täljaren avses att beskriva alla förekomster av unika ord, taggar eller liknande. Övriga angivelser av karaktären 'i-1' eller 'i-2' avser att beskriva ett förhållande till ett specifikt ord i en sekvens där ordets nummer anges som 'i'. Därför betyder 'i-1' ordet före 'i' och 'i-2' två ord före 'i'. All användning av sannolikhet bygger på frekvensen för de enskilda trigram-typerna i denna studie.

Vid disambigueringen utförs analyserna i turordning från typ (1) till typ (5) där en lämplig trigramregel vanligtvis finns bland de tre första typerna (tabell 32). Deras individuella sannolikheter anges i Ekv (18) – Ekv (22).

$$P\langle T_i|W_i,(W_{i-1},W_{i+1})\rangle = \frac{|W_i,T_i,(W_{i-1},W_{i+1})|}{|W_i,(W_{i-1},W_{i+1})|}$$

Sannolikhetsformel för lexikalisk information med kontext i EPOS. Ekv (18)

Formeln i Ekv (18) utläses så att sannolikheten för att ett ord W_i ska ha taggen T_i givet orden W_{i-1}, W_{i+1} i kontexten, är summan av alla förekomster av ordet W_i med taggen T_i i orden W_{i-1} och W_{i+1} i kontexten, dividerat med summan av alla förekomster av ordet W_i med orden W_{i-1} och W_{i+1} i kontexten. Sannolikheten i Ekv (18) kallas ibland också för lexikalisk sannolikhet.

$$P\langle T_i|W_i,(T_{i-1},T_{i+1})\rangle = \frac{|W_i,T_i,(T_{i-1},T_{i+1})|}{|W_i,(T_{i-1},T_{i+1})|}$$

Sannolikhetsformel för lexikalisk-kontextuell information med kontext i EPOS. Ekv (19)

Formeln i Ekv (19) utläses så att sannolikheten för att ett ord W_i ska ha taggen T_i givet taggarna T_{i-1}, T_{i+1} i kontexten, är summan av alla förekomster av ordet W_i med taggen T_i och taggarna T_{i-1} och T_{i+1} i kontexten, dividerat med summan av alla förekomster av ordet W_i med taggarna T_{i-1} och T_{i+1} i kontexten.

$$P\langle T_i|W_i,(T_{i-1},T_{i+1})\rangle = \frac{|T_i,(T_{i-1},T_{i+1})|}{|(T_{i-1},T_{i+1})|}$$

Sannolikhetsformel för morfosyntaktisk information, med kontext i EPOS. Ekv (20)

Formeln i Ekv (5) utläses så att sannolikheten för att ett ord W_i ska få en tagg T_i givet kontexten T_{i-1} samt T_{i+1} är summan av alla förekomster av T_i med taggarna T_{i-1} och T_{i+1} i kontexten, dividerat med summan av alla förekomster av taggarna T_{i-1} och T_{i+1} i kontexten.

$$P\langle T_i|W_i\rangle = \frac{|W_i,T_i|}{|W_i|}$$

Sannolikhetsformel för ett token, utan kontext i EPOS. Ekv (21)

Formeln i Ekv (21) utläses så att sannolikheten för att ett ord W_i får en tagg T_i är summan av alla förekomster av ordet W_i med taggen T_i dividerat med summan av alla förekomster av ordet W_i.

$$P\langle T_i|W_{ei}\rangle = \frac{|W_{ei}, T_i|}{|W_{ei}|}$$

Sannolikhetsformel för suffixinformation, utan kontext i EPOS. Ekv (22)

Formeln i Ekv (22) utläses så att sannolikheten för att ett ord W_i med ändelsen e får en tagg T_i är summan av alla förekomster av ordet med ändelsen W_{ei} med taggen T_i dividerat med summan av alla förekomster av ordet W_i med ändelsen e, W_{ei}.

I kapitel 5 redovisas utvärderingen efter disambiguering med de ovan angivna formlerna.

4.2.2.7 Gränssnitt och kommunikation

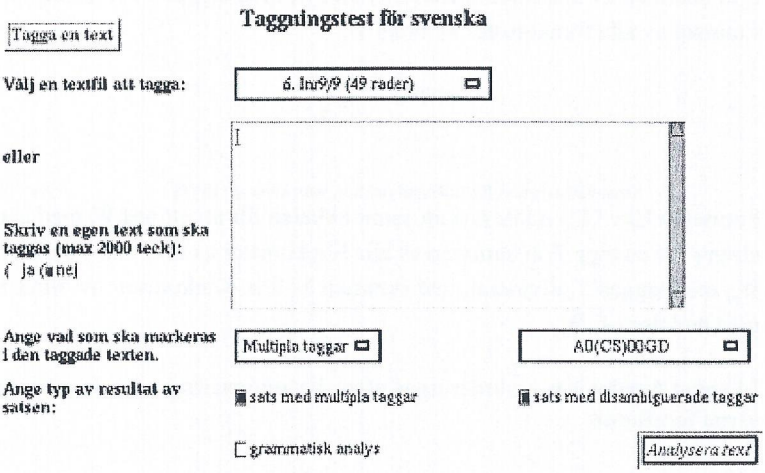
Inom ramen för avhandlingen finns det för närvarande två gränssnitt som kan användas för att studera resultatet av disambiguerad text. Det är dels ett gränssnitt utvecklat för att användas med EPOS v. 1.0, dels ett för EPOS v. 2.0. De båda gränssnitten redovisas nedan med avseende på funktion, innehåll och utseende.

Gränssnitt till EPOS 1.0

Utvecklingen av gränssnittet till EPOS 1.0 har anpassats till World Wide Web (WWW), på grund av fördelarna med plattformsoberoende. Det är dessutom relativt okomplicerat att förankra fristående verktyg till en sådan plattform. Det lämpar sig dessutom väl att använda i kombination med databasverktyget Mysql samt programmeringsspråket Perl. Som kommunikationsmedel mellan databasgränssnittet (dbi) och Perl används Common Gateway Interface (CGI).

Syftet med gränssnittet är att man ska kunna utföra olika uppgifter som är förknippade med ordklasstaggning, disambiguering och grammatisk parsning. Textmaterial som behandlas är antingen färdiga tokeniserade exempeltexter eller egen inmatad text. Därefter väljs funktion och syfte med uppmärkningen, t.ex. om man söker efter ej disambiguerade taggar, okända ord eller viss ordklass. I en analys kan ett eller flera av följande alternativ visas i en sats: multipla taggar, vilket är tillståndet före disambiguering, disambiguerade taggar samt en grammatisk analys illu-

strerad i form av en trädstruktur. Gränssnittets utseende framgår av figur 16.



Figur 16. Beskrivning av en del av gränssnittet i EPOS v.1.0.

Den analyserade texten illustreras i figur 17, där man i det specifika exemplet har valt en analys med information om satsens multipla taggar (1 a) samt dess disambiguerade version (1 b). Den partiella grammatiska parsningen med Steven Abneys SCOL-parser visas i (1 c).

En svartvit illustration visar inte att multipla taggar markeras med röd färg samt att övriga taggar har grön färg. En analys avslutas med information om antal ord, satser samt analystid per ord och sats.

1.a. Männen/NCUPND anklagas/VMN0P för/[A0PUSNI,CC,NCUSNI,RF0,S,VMIPA] att/[CI,CS] ha/[I,VAN0A,VMN0A] gömt/[AP0NSNI,VMU0A] narkotikan/NCUPND i/[NCNPNI,NCNSNI,RP0,S] skogen/NCUSND ./FE

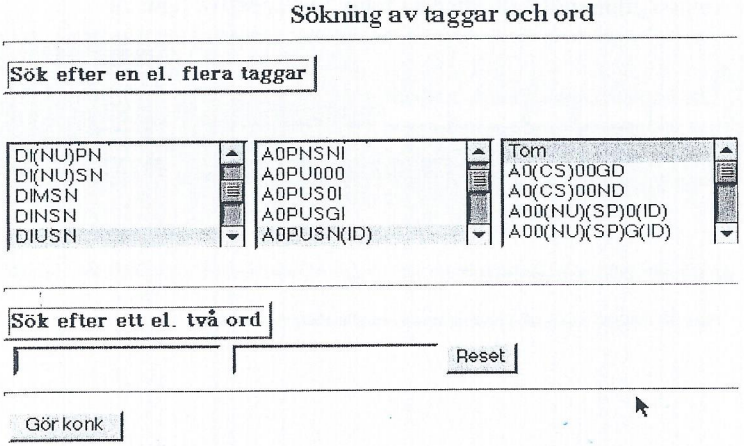
1.b. Männen/ NCUPND anklagas/ VMN0P för/ S att/ CI ha/ VMN0A gömt/ VMU0A narkotikan/ NCUPND i/ S skogen/ NCUSND ./ FE

1.c.

```
[np_ncupnd
  [NCUPND Männen]]
[vbs_op
  [VMN0P anklagas]
  [S för]]
[ATT att]
[vbs_ia
  [VMN0A ha]
  [VMU0A gömt]]
[np_ncupnd
  [NCUPND narkotikan]]
[S i]
[np_ncusnd
  [NCUSND skogen]]
[PUN .]
```

Figur 17. Beskrivning av gränssnittet efter en körning i EPOS.

I ytterligare ett gränssnitt kan man utföra sökningar på kombinationer av maximalt tre antal ord och taggar (se figur 18 och figur 19), där sökord respektive resultat illustreras.



Figur 18. Gränssnitt till sökning på ord och taggar i EPOS 1.0.

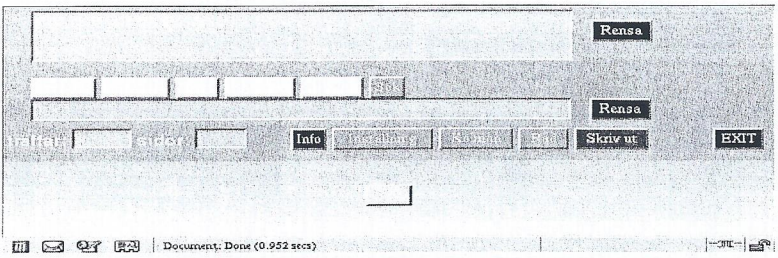
Tagg1: DIUSN	Tagg2: A0PUSN(ID)	Tagg3:	Tillbaka

MU0A flytta/VMN0A till/S	en/DIUSN egen/A0PUSN(ID) bostad/NCUSNI ./FE
rför/RJ0 utvecklat/VMU0A	en/DIUSN egen/A0PUSN(ID) femväxlad/AP0USNI låda/NC
snart/R att/CI ha/VMN0A	en/DIUSN egen/A0PUSN(ID) utrikeshandel/NCUSNI med/
en/CC slutar/VMIPA med/S	en/DIUSN egen/A0PUSN(ID) trumpetstöt/NCUSNI :/FE
SND förmånen/NCUSND av/S	en/DIUSN egen/A0PUSN(ID) chaufför/NCUSNI ./FE
esökare/NCUSNI får/VMIPA	en/DIUSN egen/A0PUSN(ID) hytt/NCUSNI som/PR(NU)(SP
yttjade/VMISA därför/RJ0	en/DIUSN egen/A0PUSN(ID) lösning/NCUSNI som/PR(NU)
	Next 20 hits

Figur 19. Gränssnitt (utvecklat inom avhandlingen) till resultat från sökning på taggar i EPOS 1.0.

Gränssnitt till EPOS v. 2.0

I en satsning på morfosyntaktisk uppmärkning av existerande textkorpusmaterial i Språkbanken utvecklas ett korpusbearbetningsverktyg. Verktyget har ett gränssnitt till World Wide Web (WWW), speciellt anpassat för sökverktyget Mozilla. Syftet med verktyget är att man ska kunna söka efter ord, taggar eller lemman i olika kombinationer genom att specificera avancerade söksträngar. Eftersom verktyget fortfarande är under utveckling är sökning idag möjlig på ett token eller en tagg åt gången, med möjlighet att välja sortering på för- och efterkontext, i textmaterial disambiguerat med EPOS 2.0[16]. Specificering av söksträng samt resultat från en sökning i gränssnittet illustreras i figur 20 respektive figur 21.



Figur 20. Sökgränssnitt till Språkbankens korpusbearbetningsverktyg.

16. Enligt personlig förmedling från Torgny Rasmark, Språkbanken, april 2002.

```
ut/PL/ och/KN/ fiska/VB INF AKT/ ./MAD/ Det/PN
för/PP/ att/IE/ fiska/VB INF AKT/ och/KN/ plocka,
PLU DEF SUB/ fiska/VB PRT AKT/ ju/AB/ ./MAD/
```

Figur 21. Resultat i sökgränssnitt till Språkbankens korpusbearbetningsverktyg.

4.2.3 Tillvägagångssätt och problem i disambigueringen

Vid ordklasstaggning måste vissa traditionella ställningstaganden göras om vilka metoder och vilka tekniker som ska användas. Dessutom kan praktiska frågor uppstå som styr val av metod och teknik. Materialets representationsform är avgörande för val av teknik; dvs den information en taggare utgår ifrån kan vara lagrad i ett speciellt format som ställer villkor på metoden som används. Andra frågor är t.ex. hur bra olika tagguppsättningar representerar den information som ska framkomma ur den taggade texten. Förutom direkt relaterade problem till ordklasstaggning fanns också problem med tekniker. För ca 4 år sedan fanns inte samma möjligheter att lika enkelt som idag tekniskt koppla ihop diverse verktyg med lexikon, och gränssnitt. Här följer en kort återblick över problem och erfarenheter från arbetet med att utveckla en taggare.

Den första frågan som gäller vilken metod som skulle användas i taggningen behandlar hur olika tekniker varierar i förmåga att hantera olika språk. Av tidsskäl begränsades arbetet till att testa endast några få representativa tekniker. Eftersom det vid tidpunkten (1997), inte fanns flera tillgängliga ordklasstaggare, utan endast Brill tagger, söktes andra möjligheter för att uppnå uppgett mål.

Initialt valdes ett manuellt angreppsätt för taggning, dels för att undersöka hur representativa manuella regler kan göras, dels för att få en uppfattning om svårigheten i att själv formulera sådana regler. Det krävs en viss kunskap om språket för att kunna avgöra i vilka sammanhang vissa taggar ska användas. Reglerna beskrev hur satsen såg ut med multipla taggar och hur de skulle transformeras till en disambiguerad motsvarighet. Endast minimal kontext runt ett element i en sats fanns med. Det fanns både mer specifika regler och mer generella. Resultatet från detta försök redogjordes för i avsnitt 4.2.2.2.

Ett andra försök var helt baserat på frekvens. Homografa ord fick sina taggar baserat på vilken tagg som var mest frekvent oavsett kontext. Jämförelsen gjordes mellan de taggar som redan hade föreslagits som multipla taggar till ett ord. Detta är utgångspunkten för en sk. *miniminivå* (*baseline*), som är maxfrekvens för varje tagg som associeras med ett token. Det är ett slags sämsta resultat som kan fås. Den skapar i normala fall runt ca 90-91% korrekt taggad text. Någon form av kontext måste dock finnas med.

Därför representerades inlärningsexempel som trigram, där en viss kontext, om än begränsad, fanns med. Detta gjordes med en fall-baserad teknik där endast generaliseringar gjordes. Sedan utökades detta med en look-

up tabell för exempel med exakt match och den fall-baserade metoden användes endast när exakt match inte kunde hittas. Denna metod används fortfarande i EPOS och kan bli ännu bättre om databasen utökas med fler exempel än de som har genererats från SUC-korpusen.

Representationsformat för optimal lagning av inlärningsmaterial var ett problem då det visade sig att det i sig utmärkta databasformat som valts i Mysql, var svårt att indexera fullständigt för maximal utsökningshastighet. Detta berodde på att olika delar av inlärningsmaterialet var ganska lika inbördes och ibland identiska, främst i den generella fall-baserade delen. För optimal indexering krävs unikt material. Detta är ett problem som bör åtgärdas i en fortsättning av avhandlingsarbetet. Tekniken fungerar tillfredsställande, men omstrukturering av inlärningsmaterialet är nödvändig för att snabba upp taggningen.

Parallellt med detta försök att använda trigram och fall-baserad metod, tränades Brill tagger på en liten del av SUC. Den visade inte särskilt bra resultat, ca 93% korrekt taggad text, och skapade grundläggande fel med t.ex. dålig kongruens i nominalfraser, vilket inte såg lovande ut. I stället för att försöka modifiera Brill tagger valdes istället EPOS för denna studie.

Ytterligare problem knutna till materialet var bl. a. koherens mellan tagguppsättningar eftersom två olika tagguppsättningar användes. Det finns starka band mellan utseendet hos det lexikaliska materialet i ett lexikon och de ordklasstaggade korpusar som används för att automatiskt extrahera ordklassmönster, när de båda källorna ska jämföras med varandra. Problem uppstår när taggarna i tagguppsättningarna inte står i ett 1:1-förhållande, dvs att alla taggar i en tagguppsättning inte har en exakt motsvarighet i den andra tagguppsättningen. Detta problem påverkar taggningen och disambigueringen genom att de meningar i en korpus där det finns problem i översättningen mellan olika tagguppsättningar inte kan användas för extrahering av regler som kan användas i disambigueringen.

När det gäller korpusevidens som disambigueringen bygger på, har det stor betydelse om utgångsmaterialet, i detta fall SUC-korpusen, är helt korrekt. Fel i korpusen orsakar i sämsta fall disambigueringsregler som är rent felaktiga. Det kan dock noteras att för varje ny version av SUC-korpusen rättas upptäckta fel till, och i den nuvarande korpusen SUC 2.0 (2002) är antal fel relativt få. Vid utvärdering av det disambiguerade materialet, är det också viktigt att den sk. 'guld-standarden' som utgörs av 10% av SUC-korpusen är korrekt, då annars korrekta analyser i jämförelsetexten som ska utvärderas tolkas som felaktiga.

Fel knutna till korpusmaterial som används för att träna en ordklasstagg-

are, väcker diskussioner om möjliga lösningar. Framför allt hur dessa fel ska hanteras med tanke på att det inte rör sig om likartade typer av fel. Det finns t.ex. taggningsfel, som måste korrigeras. Det finns andra exempel där det är tveksamt hur de ska taggas, t.ex. eliderade former som uppkommer i talspråk och inte finns med i SMDB; t.ex. "han fiska hela dan", där "fiska" (preteritum) är ett fall som inte har motsvarande morfosyntaktiska tagg i den lexikaliska databasen som ordet är taggat med i träningsmaterialet. Författarens åsikt är att sådana talspråksfall inte ska korrigeras, men de måste särbehandlas för att de inte ska kunna generera felaktiga syntaktiska disambigueringsregler.

Hantering av okända ord är ett annat problem, som visserligen analyseras och får en preliminär tagg, men vid felaktig analys orsakas följdfel i disambigueringen av homografer i närkontexten. En förbättring av denna modul är nödvändig t.ex. med speciella kontextregler som används för att avgöra om ett okänt ord mest sannolikt är ett egennamn eller ett utländskt ord. Ofta rör det sig om någon av dessa två typer av ord som är okända, ibland också nomen, men i övrigt är det få fall som involverar andra ordklasser. Detta tillhör ett av de områden som inte kan förbättras inom ramen för denna studie.

För vidare diskussioner om fel och problem i disambigueringsanalyser, se kapitel 5.

5 Utvärdering

Vid bedömning av kvaliteten i ordklasstaggning är utvärdering i någon form nödvändig. Den kan utföras på flera sätt beroende på vad som ska bedömas. Vid utvärdering av ordklasstaggning av en text är det brukligt att använda begreppet *korrekthet*. Beroende på om man vill göra en ytlig eller djupare analys i utvärderingen av taggningen är även andra mått nödvändiga.

I forskningsöversikten i avsnitt 2.2.2.3 definierades de utvärderingsmått som vanligtvis används vid utvärdering, t.ex. *precision* och *recall*. Som en vidareutveckling redovisas här två mer ovanliga utvärderingsmått eftersom det i avhandlingen också föreslås ett nytt sätt att bedöma prestation med måttet $H_{measure}$. Ett annat mått, Kappa-måttet, nämns också kortfattat. Förslaget innebär att prestation bedöms i förhållande till homografifrekvens i lexikon och homografifrekvens i text, avsnitt 5.1. I avhandlingen redovisas också delvis utförda studier och planer för fortsatt utveckling av ämnet då det verkar vara relativt outforskat för svenska språket.

I avsnitt 5.3 redovisas en jämförelse med andra utvärderingsmått av Källgren (1992) och Elworthy (1995) som tidigare angivits i forskningsöversikten.

Den metod som här har valts för utvärdering framgår av en kort beskrivning av utvärderingsverktyget (EUFOROS) i avsnitt 5.4 och av redovisningen av resultaten från utvärderingen, avsnitt 5.5. I samband med resultatredovisningen kommer olika grupper av fel att diskuteras och förslag till åtgärder för att lösa vissa problem presenteras.

5.1 Två utvärderingsmått som tar hänsyn till vissa faktorer

Som komplement till den beskrivning, av utvärderingsmått som gavs i avsnitt 2.2.2.3, redovisas här ett nytt förslag till hur utvärdering kan utföras med hjälp av $H_{measure}$. Ett annat mått, Kappa-måttet, som skiljer sig från de övriga genom att det tar hänsyn till andra faktorer som t.ex. slumpen, redovisas också kortfattat här.

Varje definition kompletteras med exempel på taggning av nomen.

5.1.1 $H_{measure}$

I syfte att komplettera de traditionella mått som används vid bedömning av analyser, föreslås här en ny typ av bedömningsmått[17], som avser att spegla homografifrekvens i stort. Den specifika homografifrekvens som här diskuteras gäller å ena sidan den i lexikonet, å andra sidan den i texter. Dels diskuteras hur själva homografifrekvensen kan mätas och dels anges ett förslag till hur den kan användas för att bedöma svårighetsgrader vid disambiguering av homografer i förhållande till just homografifrekvenser i ett lexikon respektive en text.

Anledningen till att det i detta sammanhang kan anses vara angeläget att ta fram ett nytt uttryck för att värdera prestationsmått är att SMDB som lexikalisk databas för svenska sannolikt är det lexikon som idag innehåller flest homografer. Författarens erfarenhet är att ju högre homografifrekvensen är, desto svårare är det att disambiguera homografer. Därför föreslås detta mått vilket kan användas då man vill jämföra prestation i förhållande till hur homograft ett lexikon eller en text är. En annan fördel med att ange homografifrekvens för en text är att det kan ge en indikation om textens svårighetsgrad, när det gäller just disambiguering. Det säger dock inget om andra svårigheter såsom morfologisk analys av okända ord etc. I en naturlig utvidgning av denna diskussion borde helst en fullständigare analys och testning genomföras som kan leda till konstruktion av ett fullständigare samband. Detta skulle beskriva hur svår en text är att disambiguera, vilket då skulle avspegla fler svårighetstyper än homografifrekvens. Nedan definieras två mått som kan användas för att räkna ut $H_{measure}$, H_{lex} och H_{text}. De kan även kombineras för att få ett medelvärde, H_{med}.

Formeln för att räkna ut homografifrekvensen för ett lexikon anges i Ekv (23) och formel för homografifrekvens i en text, anges i Ekv (24).

$$H_{lex} = \frac{Antal\ taggar}{Antal\ unika\ graford}$$

Formel för uträkning av homografifrekvens i ett lexikon i EPOS. Ekv (23)

$$H_{text} = \frac{Antal\ taggar}{Antal\ token}$$

Formel för uträkning av homografifrekvens i en text i EPOS. Ekv (24)

17. Måttet har granskats av Urban Hjort på avdelningen för matematisk statistik på matematiska institutionen på Chalmers.

Formeln i Ekv (25) är identisk med den traditionella 'Tag average per token', men anges ändå för förtydligande av skillnaden mellan homografimåttet i lexikon och text.

$$H_{measure} = 1 - \frac{(1 - recall)}{\left[H_{lex} | H_{text}\right]}$$

Formel för uträkning av homografins relation till mått på recall i EPOS. Ekv (25)

Sambanden i Ekv (23)-Ekv (25) kan användas för att praktiskt ta reda på homografimått i lexikon och text samt för att göra en jämförelse med hur måtten förhåller sig till reguljär användning av prestationsmått[18]. I Ekv (25) kan antingen H_{lex} eller H_{text} anges i nämnaren. Beräkning av medelvärdet av H_{lex} och H_{text}, H_{med}, visas i Ekv (30). Recall används för att ge en relation till normal prestation.

I det följande redovisas tre exempel. Det ena exemplet handlar om bedömning av svårighet i disambiguering i förhållande till homografifrekvens i lexikonet och det andra handlar om svårighet i disambiguering i förhållande till textens homografifrekvens. Det sista exemplet slår ihop måtten i de två tidigare exemplen för att ge en mer övergripande uppfattning om homografifrekvensen.

Exempel 1: Beräkning av $H_{measure}$ baserat på H_{lex}.

Antag att det finns ett lexikon, som t.ex. SMDB, vilket har 897 511 unika graford. Dessa graford har sammanlagt 1 248 680 möjliga morfosyntaktiska taggar. Se Ekv (26) för beräkning av H_{lex} (homografimåttet i lexikonet).

$$H_{lex} = \frac{Antal\ taggar}{Antal\ unika\ graford} = \frac{1248680}{897511} = 1{,}3912$$

Uträkning av homografifrekvens i lexikonet i EPOS. Ekv (26)

För att bedöma hur H_{lex} påverkar disambigueringen görs en beräkning där hänsyn tas till ett mått på recall. Måttet som tas med i denna beräkning är 94,04% vilket är ett mått uppmätt vid utvärdering av EPOS 2.0 (april 2002) på 100536 graford från SUC-materialet. Detta motsvarar ca 10% av SUC, vilken är den del som inte har använts vid träning av EPOS 2.0.

18. Jerker Järborg har medverkat i omformulering av detta samband som syftar till att möjliggöra en jämförelse med reguljära prestationsmått.

$$H_{measure} = 1 - \frac{1 - recall}{H_{lex}} = \left(1 - \frac{1 - 0,9404}{1,3912}\right) \times 100 = 95,69$$

Uträkning av $H_{measure}$ baserat på H_{lex} i EPOS. Ekv (27)

I en graf i figur 22 nedan framgår hur måttet förhåller sig till olika mått på recall med 1,39 som värde på H_{lex}.

Exempel 2: Beräkning av $H_{measure}$ baserat på H_{text}.

Antag att det finns en text som är disambiguerad, t.ex. den text som använts vid utvärderingen. Recall var 94,04%. Antal tokens i texten är 100536. Före disambiguering var det totala antalet taggar för dessa textord sammanlagt 215050. Vid uträkning av H_{text}, Ekv (28), anges ett mått på hur stor homografifrekvens som finns i texten. Detta mått motsvarar 'Tag average/token' i (Ekv 25).

$$H_{text} = \frac{Antal\ taggar}{Antal\ token} = \frac{215050}{100536} = 2,1390$$

Uträkning av homografifrekvens i texten i EPOS. Ekv (28)

För att bedöma hur H_{text} påverkar disambigueringen, görs en beräkning där hänsyn tas till ett mått på recall. Måttet som tas med i denna beräkning är, som ovan, 94,04%.

$$H_{measure} = 1 - \frac{1 - recall}{H_{text}} = \left(1 - \frac{1 - 0,9404}{2,1390}\right) \times 100 = 97,67$$

Uträkning av $H_{measure}$ baserat på H_{text} i EPOS. Ekv (29)

För att illustrera hur måttet förhåller sig till olika mått på recall med samma H_{text}, visas resultatet i en graf nedan, figur 22.

Exempel 3: Beräkning av $H_{measure}$ för både H_{lex} och H_{text}, där ett medelvärde skapas av H_{lex} och H_{text}, H_{med}.

För att ta med beräkningar om H_{lex} och H_{text} i $H_{measure}$, enligt exempelberäkningar i exempel (1) och exempel (2) ovan, kan man använda sig av följande uträkning dels för medelvärde av H_{lex} och H_{text}, Ekv (30).

$$H_{med} = \left(\frac{H_{lex} + H_{text}}{2}\right) = \left(\frac{1,3912 + 2,1390}{2}\right) = 1,7651$$

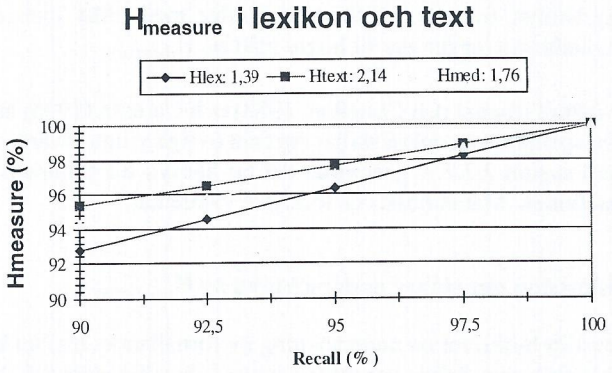
Uträkning av medelvärde för H_{lex} och H_{text} i EPOS. Ekv (30)

Medelvärdet på homografimåtten används sedan för beräkning av $H_{measure}$.

$$H_{measure} = 1 - \frac{1 - recall}{H_{med}} = \left(1 - \frac{1 - 0,9404}{1,7651}\right) \times 100 = 96,62$$

Uträkning av $H_{measure}$ baserat på H_{med} i EPOS. Ekv (31)

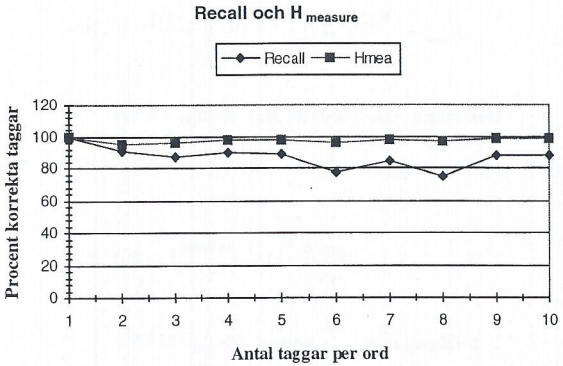
Även måttet med både H_{lex} och H_{text} och medeltalet H_{med}, visas i en graf, figur 22, för att klargöra hur måttet skiljer sig beroende på recall.



Figur 22. Förhållandet mellan medelvärdet av $H_{measure}$ i lexikon (Hl_{ex}) och text (H_{text}), med värden från EPOS 2.0.

Som framgår av figur 22 minskar homografmåttens betydelse för $H_{measure}$ ju högre värde på recall som anges. I grafen ges exempel på olika värden på recall för att illustrera hur $H_{measure}$ varierar.

I figur 23 nedan visas förhållandet mellan antal korrekta taggar och antal taggar per ord.



Figur 23. $H_{measure}$ och recall med procent korrekta taggar och antal taggar per ord i EPOS.

Av diagrammet framgår att för homografer med olika många möjliga morfosyntaktiska taggar ges ett högre mått på $H_{measure}$.

Kappa-måttet (Siegel och Castellan, 1988) och Carletta (1996) är ett annat utvärderingsmått som också skiljer sig från övriga ytliga utvärderingsmått, angivna i avsnitt 2.2.2.3, genom att det tar hänsyn till slumpens inverkan på prestationen. Måttet beskrivs dock inte vidare här.

5.2 Diskussion om vidare undersökning av $H_{measure}$

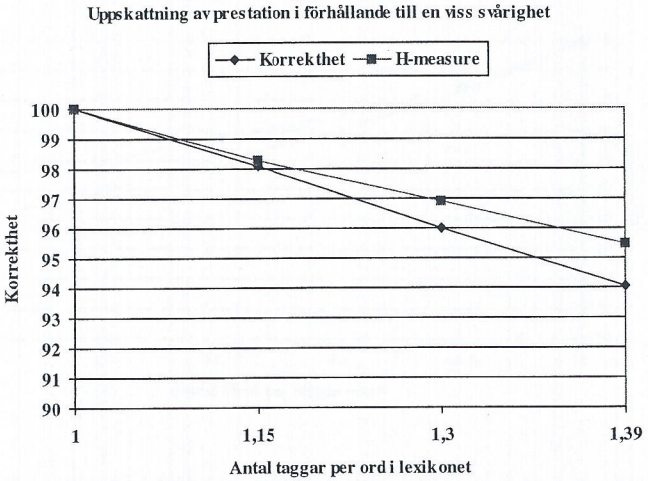
Detta avsnitt beskriver en undersökning av förhållandet mellan homografi och prestation som delvis har utförts inom avhandlingsarbetet. Nedan anges hur en sådan studie kan genomföras med hjälp av utvärderingsmåtten $H_{measure}$ respektive för jämförelse *korrekthet* (recall). Med förekomst av homografi avses både den som förekommer i lexikonet och den som förekommer före disambiguering i en text.

Syftet med måttet $H_{measure}$ är:

> Att värdera prestationen i förhållande till en viss svårighet (i detta fallet fler taggar per ord). Dvs. ju fler taggar per ord desto högre värderas prestationen.

I figur 24 nedan illustreras hur prestationen förhåller sig till en viss svårighet, dvs. svårigheten att disambiguera homografa ord. I detta fall anges

homografin som finns i lexikonet, dvs. 1,39 taggar/ord. Som framgår skiljer sig kurvan som $H_{measure}$ anger från den som anger korrekthet. Detta för att $H_{measure}$ uppvärderar prestationen. Resultatet av att använda $H_{measure}$ för utvärdering är att det resulterande måttet blir högre ju svårare uppgiften är, dvs. ju högre homografin är.



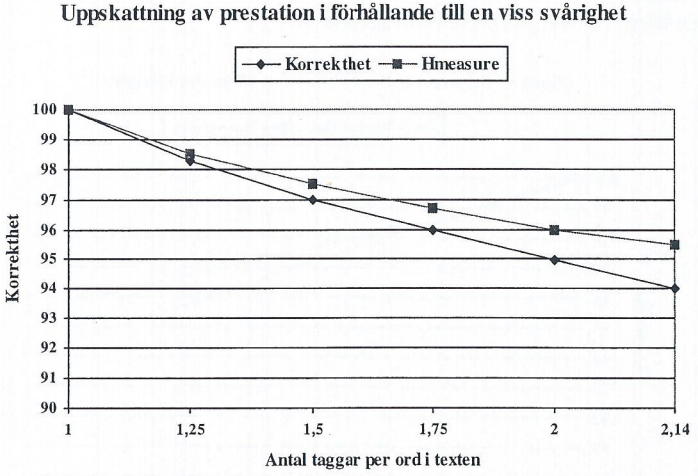
Figur 24. Uppskattning av prestation i förhållande till en viss svårighet. Homografi i lexikonet. Från EPOS.

Målet med $H_{measure}$ är alltså att beskriva ett nytt värde på prestationen (korrektheten) som stiger i förhållande till svårigheten. Minsta värde bör vara den angivna korrektheten.

Författarens idé med att använda ett nytt mått, $H_{measure}$, är att visa att korrektheten sjunker ju högre antal taggar per ord som förekommer. Detta vill författaren bevisa genom att hitta texter med olika genomsnitt av antalet taggar/ord. Antingen genom att jämföra olika textgenrer eller genom att slumpvis välja ut texter och undersöka deras genomsnittliga mått på taggar/ord.

De 100.000 ord som ingår i utvärderingen har ett genomsnitt på 2,14 taggar/ord. Därför kan det endast bevisas vilken prestation som förekommer vid två tillfällen i en text, se figur 25, dels då texten endast innehåller 1 tagg/ord, då prestationen blir 100%, och dels då texten innehåller 2,14 taggar/ord, då prestationen blir 94,04%. Värdena på prestation med

användning av $H_{measure}$ blir givetvis högre än de för korrektheten. Dessutom stiger värdena med $H_{measure}$ i takt med svårighetsgraden, dvs. ju högre homografin i texten är.



Figur 25. Uppskattning av prestation i förhållande till en viss svårighet. Homografi i texten. Från EPOS.

Slutsatserna av en studie som denna visar på det faktum att homografin är olika stor i olika typer av texter och textgenrer. Detta har undersökts för engelska texter, men det är ett relativt outforskat område för svenska texter. Det intressanta med att utföra en sådan undersökning fullt ut är att få reda på: på vilket sätt och hur mycket sådan homografi påverkar prestationen i ordklasstaggning. En fortsättning av denna studie tas upp i avsnittet om slutsatser, som en av punkterna i förslag på vidare forskning.

Slutsatserna som kan dras med de uppgifter som finns tillgängliga nu utan fördjupande studier är att homografi varierar storleksmässigt mellan texter och textgenrer.

En text med ett visst genomsnitt på taggar/ord säger endast en viss del om svårigheten att tagga eller disambiguera. Men svårigheten att disambiguera beror också på vilka typer av ord som en text innehåller och inte enbart på hur många möjliga taggar en homograf kan ha utan snarare på hur många belägg ordet har i den träningstext som används vid skapandet av disambigueringsverktyget. Antal och kvalitet på belägg avgör hur många och hur säkra disambigueringsreglerna i taggningsverktyget sedan blir.

5.3 Jämförelse med andra utvärderingsmått

I forskningsöversikten nämns två andra sätt att värdera en taggares prestation av Källgren (1992) och Elworthy (1995).

Källgren delar upp identifierade fel i kategorifel och morfologiska fel. Enligt Källgren finns det vid utvärdering av taggning av SUC betydligt fler kategorifel än morfologiska fel, men, som framgår av tabell 33, blir fördelningen mellan feltyper vid taggning med SMDB och EPOS jämn. Om några fel ska anses vara grövre än andra, bör det vara just den del som innehåller kategorifel.

Mått	Procent
Korrekthet	94,04 %
Kategorifel	51,10 %
Morfologiska fel	48,9 %

Tabell 33. Kategorifel och morfologiska fel i EPOS 2.0

Elworthy tar å andra sidan upp prestation fördelat mellan kända och okända ord. Utvärderingen som gjorts har samma fördelning i separata avsnitt. Se vidare i avsnitt 5.5.

5.4 Utvärderingsverktyg

Det verktyg som har använts för att beräkna *recall*, *precision* och *F-measure* har utvecklats inom avhandlingsprojektet. Det kallas EUFOROS (Ett Utvärderingsverktyg För Ordklasstaggning och gRammatisk parsning Och felanalyS). Utvärderingsverktyget är tänkt att kunna användas för utvärdering av ordklasstaggad text, grammatisk parsad text och utvärdering av grammatisk felsökning. Formatet på materialet som utvärderas följer EAGLES riktlinjer. Modellen i EUFOROS, vilken utvärderingen av ordklasstaggad text bygger på, illustreras i figur 26.

Verktyget och dess ingående moduler beskrivs stegvis (1-5):

1) *Indata*: Av den text som ska utvärderas finns flera versioner. Det är (a) lookup-märkt text, (b) disambiguerad text och (c) gold standard-text. Text (a) är den som innehåller alla möjliga taggar för varje token, dvs indata till disambigueringen. Text (b) är resultatet av disambigueringen. Text (c) är den 'korrekta' version av texten som ska användas som facit vid bedömning av text (b).

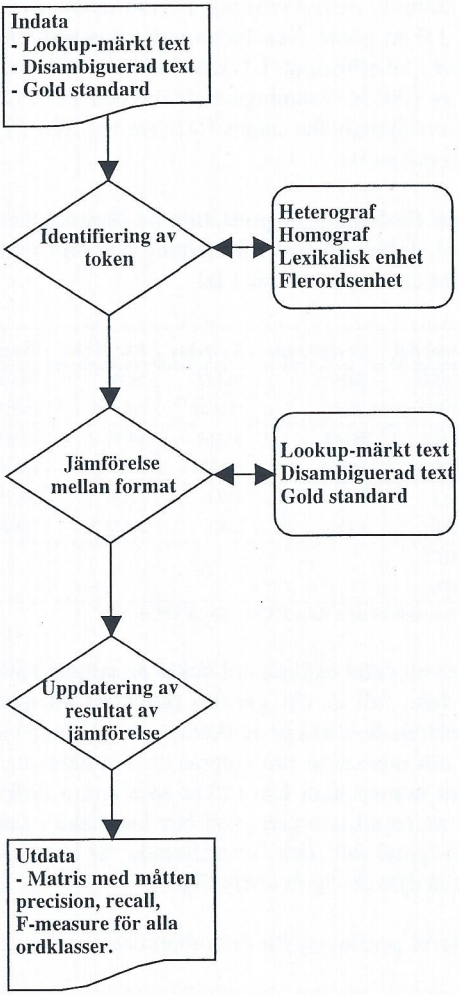
Indata kontrolleras så att texterna innehåller samma antal satser och ord.

2) *Identifiering av token*: För varje token i texten jämförs data från de tre olika texterna, men först måste varje token och dess motsvarigheter i alla texter identifieras. Identifieringen sker dels för att särskilja heterografer från homografer och dels lexikaliska enheter från flerordsenheter. En speciell svårighet utgörs av flerordsenheterna eftersom de inte är identiska i de olika texterna. Speciellt gäller detta egennamn, som har identifierats i SUC som flerordsenheter. Med SMDB:s uppmärkning identifieras de som enskilda lexikaliska enheter i de flesta fall. Flerordsenheter i form av egennamn har jämförts i utvärderingen, men övriga flerordsenheter som t.ex. adverbet 'i morgon' och 'i', 'morgon', har inte kunnat jämföras då en sådan analys kräver en speciell lista med liknande flerordsenheter och deras motsvarigheter.

3) *Jämförelse mellan format*: Här sker jämförelsen mellan de tre texterna, då alla teckensekvenser har identifierats. Först jämförs den lookup-märkta texten (a) med gold standard-texten (c) för att undersöka om den korrekta taggen för ett token finns med bland möjliga taggar i text (a). Om detta inte är fallet, t.ex. ordet 'andra', som i SUC har en adjektivtagg och i SMDB saknas detta alternativ rapporteras ett 'lookup-fel' som anger att den korrekta taggen i facit inte finns i texten. Detta görs för att skilja fel som detta från disambigueringsfel, då felet består i att fel tagg har valts. Andra fall då korrekt tagg finns i text (a) bedöms sedan genom att aktuellt token i text (b) jämförs med samma token i text (c). Resultaten från jämförelserna rapporteras i steg (4).

4) *Uppdatering av resultat av jämförelse*: Här registreras olika typer av information som framkommer från jämförelsen i steg (3). Det gäller t.ex. alla korrekta analyser, felaktiga analyser och alla 'lookup-fel'. Här finns även information om vilken analys som har använts vid disambiguering och morfologisk analys av okända ord. Resultaten rapporteras i steg (5).

5) *Utdata*: I utdata rapporteras en matris med precision, recall och F-measure för alla ordklasser, vilka ord som har analyserats korrekt, fel och vilka analyser som har använts i disambigueringen för analyserna i både de korrekta och de felaktiga fallen. Informationen i utdata är mycket användbar vid förbättring av disambigueringen. Resultaten redovisas i avsnitt 5.5.



Figur 26. EUFOROS och modulerna i utvärderingsverktyget utvecklat inom avhandlingen.

Sammanfattningsvis kan konstateras att det för utvärdering av större text-material krävs ett verktyg som EUFOROS då samma arbete skulle vara alldeles för tidskrävande att utföra manuellt.

5.5 Resultat av olika utvärderingar

I detta avsnitt kommer en redovisning av resultat av utvärdering av EPOS 2.0 och EPOS 1.0 att göras. Resultaten redovisas som en sammanfattning av flera analyser i tabellformat. Utvärdering av EPOS 2.0 (april 2002) bestod av analys av 100536 disambiguerade ord och av EPOS 1.0 (juni 2000) av 9190 ord. I resultattabeller anges först värden från EPOS 2.0 och från EPOS 1.0 inom parentes.

I tabell 34 anges fördelning av prestation på disambiguerade ord, okända ord och antal fel. Tabell 34 anger fördelning och prestation på ord, disambiguerade ord, okända ord och antal fel.

Grupper	Antal fall	Identifierade	Korrekta	Recall (%)	Precision (%)	F-measure (%)
Antal ord	100536 (9187)	100533 (9183)	94542 (8862)	94,04 (96,46)	94,04 (96,50)	94,04 (96,48)
Antal dis.	98347 (9183)	98347 (9183)	93264 (8862)	94,83 (96,46)	94,83 (96,46)	94,83 (96,46)
Okända ord	2112 (642)	2112 (642)	1114 (618)	52,75 (96,0)	52,75 (96,0)	52,75 (96,0)
Antal fel	5910 (325)					

Tabell 34. Prestationsinformation från EPOS 2.0 och EPOS 1.0.

I tabell 34 anges att antal okända ord också är samma antal som EPOS har identifierat. I detta fall är det ganska självklart att systemet hittar alla okända ord eftersom de också är markerade som sådana i indata. Men i följande tabeller, där prestation presenteras över ordklasser, identifieras inte alla nomen som nomen utan kan tolkas som andra ordklasser. I tabellen framgår också att recall och precision blir identiska i samtliga kategorier på grund av att 'antal fall' och 'identifierade' är identiska i två av fallen och i det tredje skiljer de sig åt obetydligt.

I tabell 35 beskrivs prestation för disambiguering av homografer.

Grupper	Antal	Procent
Antal disambiguerade ord	98347 (9183)	97,82 (99,9) % av alla ord
Korrekt disambiguerade ord	93264 (8862)	92,77 (96,5) % av alla ord 94,83 (96,5) % av alla dis. ord
Disambigueringsfel	5083 (325)	5,05 (3,5) % av alla ord 3,87 (3,5) % av alla dis. ord

Tabell 35. Information om disambiguering i EPOS 2.0 och 1.0.

I tabell 36 beskrivs prestation för morfologisk analys, dvs. analys av

okända ord.

Grupper	Antal	Procent
Antal okända ord	2112 (642)	2,1 (7,0) % av alla ord
Korrekt gissade okända ord	1285 (618)	1,28 (6,7) % av alla ord 60,84 (96) % av alla okända ord
Gissningsfel	827 (24)	0,82 (0,26) % av alla ord 39,16 (4) % av alla okända ord

Tabell 36. Information om analys av okända ord i EPOS 2.0 och 1.0.

I tabell 37 och tabell 38 anges disambigueringsfel[19] och gissningsfel. Dessa fel anges tillsammans i tabell 38.

Grupper	Antal	Procent
Antal fel	5910 (349)	5,87 (3,8)% av alla ord
Gissningsfel	827 (24)	0,82 (0,26)% av alla ord 13,99 (6,9)% av alla fel
Disambigueringsfel	5083 (325)	92,77 (3,5)% av alla ord 86,01 (93,1)% av alla fel

Tabell 37. Prestation med avseende på felindelning i EPOS 2.0 och 1.0.

Tidigare har sammanlagda prestationsresultat angivits, i följande avsnitt kommer en mer preciserad fördelning att göras med avseende på ordklasser.

5.5.1 Prestation indelad i ordklasser

I detta avsnitt ges en närmare beskrivning av hur prestationen är fördelad mellan olika ordklasser, dels efter disambigueringen och dels efter analysen av okända ord.

Eftersom den sammanlagda prestationen vid disambiguering inte säger något om hur prestationen fördelas mellan olika ordklasser, beskrivs detta i tabell 38. Måtten recall, precision och F-measure anges i procent.

19. Av de disambigueringsfel som anges ingår ca 12% av felen som kan kategoriseras som "lookup-fel". Dessa fel beror inte på fel i disambigueringen utan snarare på annoteringsdelen som föregår disambigueringen.

Grupper	Antal fall	Identifierade	Korrekta	Rec.	Prec.	F-mea.
Nomen	20347	20555	19113	93,94	92,98	93,46
	(1867)	(1873)	(1786)	(95,66)	(95,35)	(95,51)
Egennamn	3931	3817	3131	79,65	82,03	80,94
	(698)	(683)	(665)	(95,27)	(97,36)	(96,32)
Verb	15650	15529	15031	96,04	96,79	96,42
	(1397)	(1396)	(1374)	(98,35)	(98,42)	(98,39)
Adjektiv	6832	6500	5775	84,53	88,85	86,69
	(646)	(645)	(590)	(91,33)	(91,47)	(91,4)
Adverb	8548	9043	7883	92,22	87,17	89,7
	(590)	(597)	(565)	(95,76)	(94,63)	(95,2)
Pronomen	7562	7646	7211	95,36	94,31	94,84
	(595)	(599)	(580)	(97,47)	(96,82)	(97,15)
Räkneord	2318	2252	2141	92,36	95,07	93,72
	(243)	(236)	(234)	(96,29)	(99,15)	(97,72)
Preposition	11712	11804	11452	97,78	97,02	97,4
	(1068)	(1053)	(1040)	(97,37)	(98,76)	(98,07)
Artikel	4847	4738	4482	92,47	94,60	93,54
	(417)	(415)	(402)	(96,40)	(96,86)	(96,63)
Interpunktion	11851	11843	11819	99,73	99,80	99,76
	(959)	(959)	(959)	(100)	(100)	(100)
Konjunktion	5572	5462	5238	94,01	95,90	94,96
	(597)	(598)	(564)	(94,47)	(94,31)	(94,39)
Utländska ord	63	35	14	22,22	40	31,11
	(5)	(12)	(5)	(100)	(41,66)	(70,83)
Interjektion	57	53	48	84,21	90,57	87,39
	(2)	(5)	(2)	(100)	(40)	(70)
Förkortning	122	110	97	79,51	88,18	83,84
	(2)	(2)	(2)	(100)	(100)	(100)
Infini-	1124	1146	1107	98,49	96,60	97,54
tivmärke	(101)	(110)	(94)	(93,06)	(85,45)	(89,26)
TOTALT	100536	100533	94542	94,04	94,04	94,04
	(9187)	(9183)	(8862)	(96,46)	(96,50)	(96,48)

Tabell 38. Prestation med avseende på ordklassindelning vid disambiguering med EPOS 2.0 och 1.0.

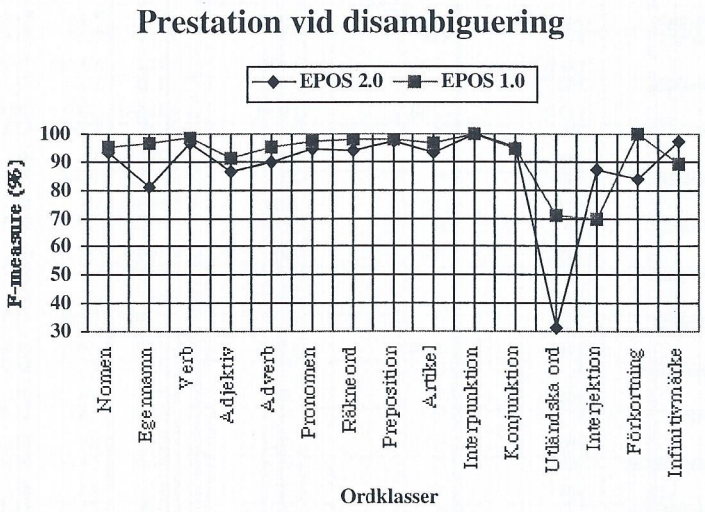
I tabell 39[20] nedan ges en motsvarande fördelning över prestation av analys av okända ord.

20. Vissa smärre frågetecken i annoteringen kvarstår vilket är en förklaring till förekomster av okända ord bl.a. i ordklasserna preposition, adverb och pronomen, vilka annars inte är högfrekventa när det gäller okända ord.

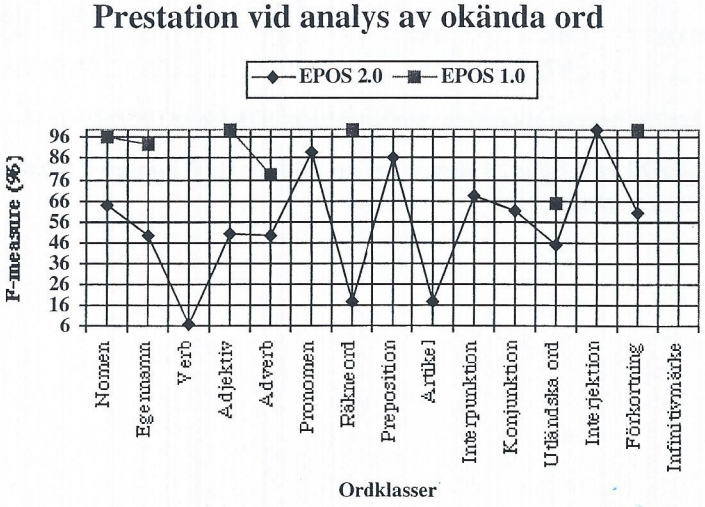
Grupper	Antal fall	Identifierade	Korrekta	Rec.	Prec.	F-mea.
Nomen	805	868	532	66,09	61,29	63,69
	(315)	(325)	(309)	(98,1)	(95,08)	(96,59)
Egennamn	941	888	454	48,25	57,27	49,76
	(127)	(109)	(109)	(85,83)	(100)	(92,92)
Verb	19	16	1	5,26	7,69	6,48
	(0)	(0)	(0)	(0)	(0)	(0)
Adjektiv	114	91	51	44,74	56,04	50,39
	(27)	(27)	(27)	(100)	100	100
Adverb	16	21	9	56,25	42,86	49,56
	(4)	(7)	(4)	(100)	(57,14)	(78,57)
Pronomen	19	19	17	89,47	89,47	89,47
	(0)	(0)	(0)	(0)	(0)	(0)
Räkneord	23	22	4	17,39	18,18	17,78
	(165)	(165)	(165)	(100)	(100)	(100)
Preposition	8	11	8	100	72,72	86,36
	(0)	(0)	(0)	(0)	(0)	(0)
Artikel	3	54	1	33,33	1,85	17,59
	(0)	(0)	(0)	(0)	(0)	(0)
Interpunktion	43	32	25	58,14	78,12	68,13
	(0)	(0)	(0)	(0)	(0)	(0)
Konjunktion	2	9	2	100	22,22	61,11
	(0)	(0)	(0)	(0)	(0)	(0)
Utländska ord	39	5	4	10,26	80	45,13
	(2)	(7)	(2)	(100)	(28,6)	(64,3)
Interjektion	5	5	5	100	100	100
	(0)	(0)	(0)	(0)	(0)	(0)
Förkortning	5	1	1	20	100	60
	(2)	(2)	(2)	(100)	(100)	(100)
TOTALT	2042	2042	1114	54,55	54,55	54,55
	(642)	(642)	(618)	(96,26)	(96,26)	(96,26)

Tabell 39. Prestation med avseende på ordklassindelning av okända ord i EPOS 2.0 och 1.0.

I figur 27 och figur 28 nedan ges en bild av fördelningen i prestation i form av punktdiagram.



Figur 27. Spridning av prestation vid disambiguering i form av punktdiagram från utvärdering av EPOS 2.0 och 1.0.



Figur 28. Spridning av prestation vid analys av okända ord i form av punktdiagram från utvärdering av EPOS 2.0 och 1.0.

5.5.2 Analys av fel

I detta avsnitt redovisas vilka typer av fel som görs i disambigueringen och analysen av okända ord.

Disambigueringsfel

För att få en bättre överblick över felen som görs i disambigueringen visas i tabell 40 olika typer av fel och deras fördelning mellan ordklasserna. Av beskrivningen framgår vilka svagheter som finns i ett system, vilket är ett mer eller mindre nödvändigt underlag för förbättring av disambigueringen. Efter tabellen följer en diskussion om fel, orsaker och åtgärder.

Ord som skulle ha blivit	Blir istället:	Antal:	Procent av alla fel
Nomen	Egennamn	23 (14)	0,45 (4,3)
	Verb	53 (1)	1,04 (0,3)
	Nomen (numerusfel)	636 (53)	12,51 (16,3)
	Adverb	45 (1)	0,88 (0,3)
	Adjektiv	35 (1)	0,68 (0,3)
	Partikel	- (1)	(0,3)
	Utländska ord	4 (3)	0,078 (0,9)
	Nomen (gen.&best.)	131 (1)	2,58 (0,3)
	Interjektion	3	0,06
	Räkneord	5	0,10
	Konjunktion	2	0,04
	Förkortning	5	0,10
	Pronomen	13	0,25
	Nomen (övriga)	6 (5)	0,12 (1,5)
Summa Nomen		961 (80)	18,91 (24,5)
Pronomen	Verb	9 (1)	0,18 (0,3)
	Artikel	123 (6)	2,42 (1,8)
	Konj.	71 (3)	1,40 (0,9)
	Adverb	63 (4)	1,24 (1,2)
	Adjektiv	2 (1)	0,04 (0,3)
	Nomen	42	0,82
	Räkneord	14	0,28
	Övriga	25	0,49
Summa Pronomen		349 (15)	6,87 (4,6)
Adjektiv	Nomen	78 (3)	1,53 (0,9)
	Verb	109 (12)	2,14 (3,7)
	Adj. (num&best.)	136 (19)	2,68 (5,8)
	Pronomen	2 (1)	0,04 (0,3)
	Adverb	418 (6)	8,22 (1,8)
	Egennamn	7 (2)	0,13 (0,6)
	Adj. (particip)	161 (7)	3,16 (2,1)

	Räkneord	(1)	(0,3)
	Konjunktion	1	0,02
	Adj. (övriga)	66 (5)	1,30 (1,5)
Summa Adjektiv		978 (56)	19,24 (17,2)
Verb	Verb (tempusfel)	184 (7)	3,62 (2,1)
	Adjektiv	67 (10)	1,31 (3,0)
	Nomen	40 (3)	0,78 (0,9)
	Adverb	206 (2)	4,05 (0,6)
	Preposition	5 (1)	0,10 (0,3)
	Egennamn	1	0,02
	Pronomen	6	0,12
	Utländska ord	4	0,08
	Verb (passiv/aktiv)	70	1,38
	Övriga	8	0,16
Summa Verb		591 (23)	11,62 (7,0)
Preposition	Adverb	200 (3)	3,93 (0,9)
	Verb	5 (1)	0,10 (0,3)
	Konjunktion	47 (9)	0,92 (2,7)
	Adverbialpartikel	(13)	(4)
	Egennamn	(2)	(0,6)
	Utländska ord	5	0,10
	Adjektiv	1	0,02
	Nomen	2	0,04
Summa Preposition		260 (28)	5,12 (8,6)
Artikel	Pronomen	258 (13)	5,08 (4)
	Räkneord	30 (1)	0,60 (0,3)
	Adverb	47 (1)	0,92 (0,3)
	Nomen	6	0,12
	Egennamn	1	0,02
	Verb	4	0,08
	Utländska ord	1	0,02
	Artikel (numerus)	5	0,10
	Övriga	11	0,22
Summa Artikel		363 (15)	7,14 (4,6)
Adverb	Adjektiv	198 (10)	3,90 (3,0)
	Preposition	254 (9)	5,0 (2,7)
	Konjunktion	67 (3)	1,32 (0,9)
	Pronomen	28 (1)	0,55 (0,3)
	Adverbialpartikel	(2)	(0,6)
	Artikel	1	0,02
	Nomen	32	0,63
	Adverb (komparation)	51	1,0
	Verb	25	0,49
	Utländska ord	2	0,04
Summa Adverb		658 (25)	12,94 (7,7)

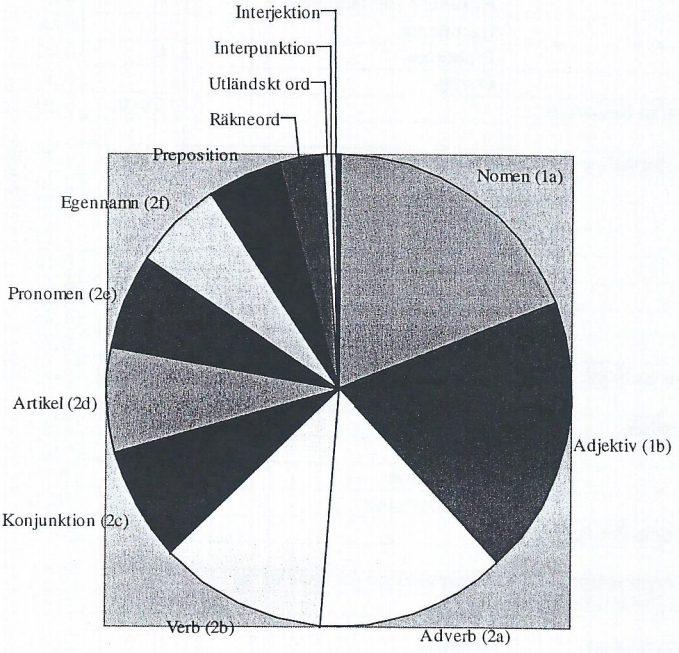
Egennamn	Nomen	86 (21)	1,69 (6,4)
	Adjektiv	4 (1)	0,08 (0,3)
	Utländska ord	(3)	(0,9)
	Interjektion	(3)	(0,9)
	Adverb	(2)	(0,6)
	Egennamn (genus)	118	2,32
	Egennamn (kasus)	92	1,80
	Förkortning	7	0,14
	Verb	1	0,02
	Övriga	4	0,08
Summa Egennamn		312 (30)	6,14 (9,2)
Räkneord	Artikel	78 (7)	1,53 (2,1)
	Nomen	7 (1)	0,14 (0,3)
	Adjektiv	(1)	(0,3)
	Adverb	1	0,02
	Utländska ord	1	0,02
	Verb	5	0,10
	Räkneord (numerus)	18	0,35
	Räkneord (ordnings-/grundtal)	2	0,04
	Räkneord (bestämdhet)	24	0,47
	Egennamn	2	0,04
	Pronomen	18	0,35
	Övriga	2	0,04
Summa Räkneord		158 (9)	3,10 (2,7)
Konjunktion	Preposition	90 (2)	1,77 (0,6)
	Adverb	115 (9)	2,26 (2,7)
	Konjunktion (Und.&sam.)	12 (18)	0,24 (5,5)
	Pronomen	65 (4)	1,28 (1,2)
	Nomen	8	0.16
	Adjektiv	2	0,04
	Utländska ord	65	1,28
	Infinitivmärke	39	0,76
Summa Konj.		396 (33)	7,79 (10,1)
Partikel	Konjunktion	(1)	(0,3)
	Adverb	(4)	(1,2)
	Preposition	(1)	(0,3)
	Utländska ord	(1)	(0,3)
Summa Partikel		(7)	(2,1)
Interpunktion	Interpunktion (intern/extern)	13	0,25
Summa Interpunktion		13	0,25
Interjektion	Nomen	8	0,16
	Egennamn	1	0,02
Summa Interjektion		9	0,18
Utländska ord	Adjektiv	1	0,02
	Nomen	21	0,41
	Egennamn	5	0,1

	Pronomen	2	0,04
	Adverb	1	0,02
	Verb	2	0,04
	Interpunktion	3	0,06
Summa Utl. ord		35	0,68
Summa Totalt		5083 (325)	

Tabell 40. Indelning av fel från disambiguering med EPOS 2 och 1. (5083 resp. 325 st.)

Nedan följer en beskrivning, diskussion och ibland även förslag till lösningar på de vanligaste disambigueringsfelen i EPOS 2.0. (Det är inte meningsfullt att föreslå åtgärder för EPOS 1.0 som inte längre används.) De två största grupperna är nomen och adjektiv, 1a respektive 1b. De fem största grupperna därefter är adverb, verb, konjunktion, artikel, pronomen och egennamn. De betecknas 2a-f.

Cirkeldiagramet beskriver fördelningen över disambigueringsfel (figur 29).



Figur 29. Cirkeldiagram över fördelning av disambigueringsfel i EPOS 2.0.

Grupp 1a och 1b

De två största grupperna av disambigueringsfel gäller felklassificerade nomen och adjektiv. I grupp *1a*, är de vanligaste felen av morfologisk karaktär men ibland tolkas nomen som egennamn.

Nomen får ofta fel numerusform. Numerusform på nomen är inget trivialt problem när man gör en ganska ytlig syntaktisk analys. Ofta är kontexten för liten för att avgöra vilken form som är rätt i ett visst sammanhang. Det bästa är att ha en mekanism som känner igen vilka pronomen, adjektiv och artiklar som hör ihop med vilka nomen. Detta kan möjligtvis hanteras på en högre nivå, i en grammatisk parsning, då kontexten blir större. Men det finns också de ord som inte har mycket information i kontext, ofta anges de i pluralform, t.ex. 'ägare', 'tränare', 'folk', 'lopp' etc. Annan morfologisk information som är problematisk är genus och bestämdhet. Exempel på sådana ord är 'pris', 'lag' där genusbestämning är ett problem och i fall som 'längtan', 'inverkan' och liknande ord är bestämdheten svår att avgöra. Av de morfologiska problemen är gruppen med numerusproblem störst. Möjliga åtgärder är t.ex. att identifiera alla nomen som frekvent används utan ledande information i kontext som pluralis obestämd form.

Problem med att särskilja nomen och egennamn uppkommer även i den motsatta riktningen, dvs. egennamn som tolkas som nomen. En lösning på detta är att ha en modul som kan identifiera egennamn. Det räcker inte att titta på om första bokstaven i ett ord är en versal, dels därför att man inte kan förutsätta att alla ord som ska vara egennamn faktiskt stavas rätt med versal, dels därför att alla ord som börjar med versal inte alltid är egennamn. En generell tolkningsregel som har använts här är att substantiviska ord med versal som kan böjas i annan form än genitiv måste tolkas som nomen och de med versal som inte kan böjas är egennamn. Detta är en subjektiv tolkning och något som man kan diskutera. Det är dessutom ett av inkonsekvensproblemen i SUC-korpusen, där nomen med stor bokstav ibland tolkas som egennamn och ibland inte.

Adjektiv, grupp *1b*, är den andra stora problemgruppen i disambigueringen. De vanligaste felen är att adjektiven förväxlas med verb eller blir adjektiv men får fel numerus eller bestämningsform. Att skilja mellan när ett ord ska tolkas som adjektiv eller verb kräver egentligen en speciell analys eftersom det är ett ganska stort problem och det är heller inte enkelt att lösa. När adjektiven får fel numerus eller bestämningsform beror det oftast på kontexten. Om det i kontexten finns ett nomen som har fel numerusform får adjektivet det också. Det finns en speciell kontroll i disambigueringsmodulen som har som uppgift att se till att främst adjektiv och nomen överensstämmer i genus, numerus och bestämning, men det

hjälper inte om felet ligger hos nomenet. Därmed kan de flesta morfologiska fel hos adjektiv kategoriseras som följdfel.

Grupp 2a-2f

Efter de två största grupperna av disambigueringsfel följer sex andra grupper som är ungefär lika stora. De utgörs av adverb, verb, konjunktioner, artiklar, pronomen och egennamn.

Adverb, grupp *2a*, innehåller ord som ibland kan tolkas som adjektiv och prepositioner. Om ett adverb tolkas som adjektiv beror det oftast på att det är svårt att se att ordet är en bestämning till ett verb eller till hela satsen. Speciellt när ordet förekommer sist i en sats kan det vara ännu svårare att avgöra bestämning därför att kontexten är mer begränsad. Kontextuell information som innehåller interpunktionstecken är utan värde eftersom den förekommer på samma sätt för nästan alla ordklasser. Exempel på sådana adverb är 'mycket' och 'svårt'. Gruppen där adverb tolkas som preposition har också att göra med att det är svårt att se vilket ord adverbet är bestämning till. Exempel på sådana adverb är 'med', 'till', 'på' och 'för'. I flera av fallen rör det sig om adverbialpartiklar som hör till ett partikelverb. Mer information om detta i lexikonet och särbehandling i analys bör ge bättre resultat. Gruppen 2a kan antagligen totalt sett förbättras genom att en speciell analys undersöker just adverb som eventuellt kan vara bestämningar till verb eller andra ordklasser. Detta borde göras med uppbyggnad av autentiska exempel eftersom man inte kommer speciellt långt med generaliseringar över ordklasser. Resultaten blir inte acceptabla beroende på brist på exakt matchande exempel i disambigueringsreglerna, vilket gör att generaliseringar används.

Verb, grupp *2b*, är en grupp där ord ofta förväxlas med adverb. Det finns också problem med morfologiska distinktioner inom gruppen verb. De ord som ibland felaktigt tolkas som adverb är t.ex. 'var', 'åt', 'samlat', 'arbetat'. Vissa verb med intern homografi, t.ex. 'tippas', 'planeras', 'tas', 'öppnas', kan få fel tempusform och fel s-form, dvs. passiv eller deponens. Vissa verbformer ska föregås av ett infinitivmärke vilket kan vara en ledtråd vid ansatser till förbättring av problemet. Om det finns ett 'att' med kanske det dessvärre tolkas som någon annan ordklass, t.ex. konjunktion. Om det hade funnits hjälpverb markerade i texten hade dessa kunnat användas som vägledning vid analys av vissa verb i preteritum. Hjälpverb markeras dock varken i SUC eller SMDB. Planer finns på att komplettera både SMDB och SUC som träningskorpus med den informationen.

Felen för konjunktioner, grupp *2c*, är dels de fall som blir adverb och dels de som förväxlas med prepositioner. Detta problem kan bli ganska allvar-

ligt i en högre analys, t.ex. grammatisk parsning då det är viktigt att hitta frasgränser. Denna procedur försvåras om konjunktionerna är inkorrekta. Konjunktioner tolkas ibland som adverb, t.ex. vid 'som', 'så', 'sedan'. Preposition är en vanlig felaktig tolkning på 'för', 'utan', 'om' i vissa kontexter då dessa ord i stället borde ha blivit konjunktioner. Sådana fall är speciellt problematiska. 'Som' behandlas i en särskild analys i EPOS 2.0.

Artiklar, grupp *2d*, tolkas ofta som pronomen. Det gäller vanliga ord som 'sådana', 'den' och 'de'. Om nominalfraser kan identifieras förbättras möjligheterna att särskilja artiklar från pronomen men det finns också undantag till antagandet att alla ord som kan vara artiklar i nominalfraser är det, då de trots utseendet i kontexten är pronomen och inte artiklar.

Grupp *2e*, pronomen, förväxlas främst med artiklar och konjunktioner. Artiklar som 'en', 'annan' och 'någon' är omvända feltolkningar mot dem i grupp 2d. Samma resonemang gäller därför här. Grupp 2d och 2e är alltså ett exempel på två ordklasser som ömsesidigt är svåra att urskilja. Vid felaktig tolkning av pronomen som konjunktion gäller det till största delen ordet 'som'. En särskild analys tolkas ordet 'som' men eftersom det är ett mycket frekvent ord blir felen också fler.

Egennamn, grupp 2f, är en grupp där intern homografi orsakar felaktiga val av taggar för genus- och kasusattributen i taggarna. Exempel på ord med fel genusform är 'TT-Reuter', 'Jocke', 'Leconte' där utrum borde ha angivits i stället för neutrum. Kasusform på taggar blir också frekvent fel, t.ex. 'Europas', 'Anders', 'Spartacus'. En namnigenkänningsmekanism bör kunna ge partiella lösningar på detta problem.

Slutsatser om disambigueringsfel

Slutsatserna av diskussionen om disambigueringsfel är att felen antagligen är ganska representativa då de ofta utgör lika stora grupper vid utvärdering av både EPOS 1.0 och EPOS 2.0. Det har framkommit att det finns vissa svårigheter att särskilja homografer som ömsesidigt tolkas fel, dvs. A tolkas som B och B tolkas som A. För sådana homografer är det naturligtvis mest angeläget att skapa speciella analyser. Andra analyser som är viktiga för förbättring av disambigueringen är att beskriva skillnaden mellan numerusformer i nomen, nomen och egennamn, adverb och adjektiv etc. Listan kan göras lång, och det är svårt att veta när en analys inte kan göras bättre, p.g.a. att man alltid kan pröva en annan teknik, men om man håller sig till samma metod måste man vara medveten om att när man skapar en ny modul som ska arbeta med ett nytt problemområde inom t.ex. disambiguering, skapar modulen som resultat ofta nya problem, dvs. nya fel.

Som sammanfattning av disambigueringsfelen presenteras i tabell 41 en lista över de regler som genererar mest fel.

Frekvens	Regeltyp
152	Tokenbaserad
181	Suffixbaserad
244	Lexikalisk
607	Lookup-fel
768	Bigram
904	Lexikalisk-kontextuell
2205	Kontextuell

Tabell 41. De 7 vanligaste regeltyperna vid disambigueringsfel i EPOS.

En motsvarande lista över mest korrekt använda regler presenteras i tabell 42.

Frekvens	Regeltyp
722	Infinitivmärkesanalys
740	Analys av 'som'
2622	Kongruens
2972	Lexikalisk
5682	Kontextuell
10260	Lexikalisk-kontextuell
22591	Bigram

Tabell 42. De 7 vanligaste regeltyperna vid korrekta analyser i EPOS.

De två tabellerna över regler som används vid disambiguering visar att vissa regler som frekvent blir fel, också frekvent blir korrekt som t.ex. i fallet med kontextuella regeltyper. Men regeltypen bigram har gett störst utdelning då den endast genererar ca 3% fel. På andra plats kom de lexikaliska reglerna med 7,6% fel och på tredje plats de lexikalisk-kontextuella reglerna med endast 8% fel. Sämst resultat fick den kontextuella regeltypen med 39% fel.

Fel i morfologisk analys

I tabell 43 visas en fördelning över felen vid analys av okända ord.

Ord som skulle ha blivit:	Blir istället:	Antal:	Procent av alla fel
Nomen	Nomen (numerusfel)	19 (2)	2,04 (8,3)
	Adverb	4 (1)	0,43 (4,1)
	Utländskt ord	(3)	(12,5)
	Nomen (genus)	12	1,29
	Nomen (bestämdhet)	6	0,64
	Nomen (kasus)	1	0,10

	Egennamn	164	17,65
	Adverb	4	0,43
	Konjunktion	7	0,75
	Adjektiv	10	10,76
	Räkneord	1	0,10
	Verb	3	0,32
	Artikel	46	4,95
Summa Nomen		277 (6)	29,82 (25)
Egennamn	Nomen	291 (14)	31,32 (58,3)
	Adverb	2 (2)	0,20 (8,3)
	Utländska ord	1 (2)	0,10 (8,3)
	Egennamn (genus)	134	14,42
	Egennamn (kasus)	33	3,55
	Verb	7	0,75
	Adjektiv	8	0,86
	Artikel	6	0,64
	Räkneord	3	0,32
	Pronomen	1	0,10
Summa Egennamn		486 (18)	52,31 (75)
Adjektiv	Adjektiv (numerus)	17	1,82
	Adjektiv (particip)	4	0,43
	Verb	4	0,43
	Egennamn	32	3,44
	Nomen	2	0,20
	Adverb	4	0,43
Summa Adjektiv		63	6,78
Artikel	Egennamn	1	0,10
Summa Artikel		1	0,10
Interpunktion	Egennamn	11	1,18
	Interpunktion (intern/extern)	7	0,75
Summa Interpunktion		18	1,94
Räkneord	Räkneord (numerus)	14	1,50
	Egennamn	5	0,54
Summa Räkneord		19	2,04
Pronomen	Artikel	1	0,10
	Egennamn	1	0,10
Summa Pronomen		2	0,20
Adverb	Nomen	2	0,20
	Egennamn	2	0,20
	Preposition	3	0,32
Summa Adverb		7	0,75
Verb	Verb (tempus)	1	0,10
	Adverb	2	0,20

	Egennamn	15	1,61
Summa Verb		18	1,94
Utländska ord	Nomen	3	0,32
	Egennamn	31	3,34
	Pronomen	1	0,10
Summa Utl. ord		35	3,76
Förkortning	Egennamn	3	0,32
Summa Förk.		3	0,32
Summa Totalt		929 (24)	

Tabell 43. Indelning av fel från analys av okända ord med EPOS 2 och 1. (929 resp. 24 st.)

Fel som uppstår vid analys av okända ord är relativt få och ganska begränsade. De består dels av nomen som blir egennamn, grupp 3a. De består dessutom av egennamn som istället blir nomen och egennamn med genusfel, grupp 3b.

Grupp 3a

Av de nomen som blir egennamn är några exempel 'korvkö', 'eltejp', 'teknis' och 'Viggenplan'. De är ord som inte finns i SMDB. Den morfologiska analysen har inte resulterat i något passande nomen och därför blir de egennamn. Andra ord som har tolkats som okända men är relativt lätt att åtgärda är vanliga ord skrivna med versaler. Exempel på sådana är 'ISHOCKEY' och 'JUBILEUM'. Dessa finns i SMDB, men inte skrivna med versaler.

Grupp 3b

Den andra gruppen vid analys av okända ord är felklassificerade egennamn. De tolkas istället som nomen eller får fel genusform som egennamn. Av de ord som blir nomen finns t.ex. 'Kulturhuset', 'Exportrådet', 'Systemet', 'Badrocken' och 'Ängbyskolan' etc. Det är egennamn som inte finns i SMDB men som vid morfologisk analys snarare ser ut som nomen. Ofta rör det sig om satsinitiala ord. Vid morfologisk analys undersöks i första steget om ett ord kan vara någon annan ordklass än ett egennamn; dvs att om ordet börjar med versal har det då ingen betydelse. Anledningen till detta är naturligtvis att orden alltid börjar med versal i början av en sats. Det är givetvis också en tolkningsfråga vilka ord som ska tolkas som egennamn och vilka som ska tolkas som nomen. Den andra delen egennamn som frekvent blir fel handlar om felaktiga gissningar på ord som t.ex. 'Agassi', 'Shamir' etc. Svårigheten ligger i att avgöra om det rör sig om personnamn då utrum-formen i egennamnet ska användas. Om det i

stället handlar om platser, städer, länder eller föreningar etc. skall 'neutrum-formen' av egennamnet användas. Översättning med genusinformation har utarbetats som förslag i den samarbetsgrupp som har konsulterats i avhandlingsarbetet. Syftet med att ta med genusinformation är att underlätta kongruensanalyser i nominalfraser med egennamn. Förbättringar i denna analys kan göras genom en speciell namnigenkänningsmekanism, vilket diskuterades vid disambigueringsfel av samma typ.

Övriga grupper som kan orsaka problem är förkortningar och utländska ord. Andra grupper är oftast inte aktuella vid upptäckt av nya ord såvida de inte förekommer som stavningsvarianter eller felstavningar, men den typen av fel uppkommer nästan aldrig i texter som används i utvärdering. I den text som har använts i denna utvärdering hittades två stavningsvarianter, där "ayatollahns" stavas med ett "h". I SAOL stavas det "ayatollans" utan "h". Därmed finns det en möjlighet att andra typer av liknande karaktär uppstår när ett system som EPOS testas på andra icke-kontrollerade texter.

Slutsatser om morfologisk analys

Slutsatsen man kan dra av analys av okända ord är att den främst kan förbättras om det sätts in extra resurser som t.ex. en analys som känner igen egennamn och inte förväxlar dem med nomen.

5.6 Slutsatser

De slutsatser man kan dra av den utvärdering som har gjorts här är flera. Dels är det författarens personliga erfarenhet att utvärderingsresultat alltid är starkt beroende av det material som utvärderas. Om materialet hade innehållit andra typer av texter hade resultatet kanske sett annorlunda ut, främst vad det gäller analysen av okända ord. I disambigueringsdelen finns vissa stora grupper som skiljer sig från de mindre och det framgår tydligt var insatser är nödvändiga för att förbättra disambigueringen.

Skillnaden i grupperingar av felaktiga analyser mellan EPOS 1.0 och EPOS 2.0 är förvånansvärt liten, vilket tyder på att även utvärdering på mindre material kan vara mycket representativa. Felen som rapporterats från de båda systemen handlar om samma grupper även om de inte är lika stora. Betydelsen av storlek på tagguppsättning är till största delen avgörande i skillnad i prestation mellan v. 1.0 och v. 2.0. V. 1.0 innehåller en mycket större och mer detaljerad tagguppsättning vilket förmodligen är en fördel. Det kan också antas att homografifrekvensen i SMDB har mycket stor betydelse för skillnad i prestation mellan de båda systemen.

Stora grupper som blir fel är dels de som är ömsesidigt svåra att urskilja, dels följdfel. Hur stor andel följdfelen utgör är svårt att avgöra. De bör dock inte överstiga 50% av alla fel, eftersom den form av trigram som används i disambigueringen är till för att hindra spridning av fel.

Orsaken till att fel uppstår är dels överensstämmelseproblem mellan SUC och SMDB, dels också vilken typ av analys som görs. Studien har visat att de kontextuella reglerna, som enbart bygger på analys av möjliga ordklasstaggar i kontexten, ger sämst resultat, 39% av analyserna blir fel. Däremot framkom att bigramreglerna gav oväntat bra resultat, endast 3% av analyserna var felaktiga.

Målet med denna utvärdering var att visa utvärderingsresultat av EPOS 2.0 och EPOS 1.0. Detta har gjorts med de traditionella utvärderingsmåtten, men också med hjälp av nya ekvationer som tar hänsyn till homografifrekvens i lexikon och text. Det går inte att dra några direkta slutsatser av de nya utvärderingsmått som föreslagits utan att ha andra lexikon att jämföra med. Andra texter är förmodligen också lättare att jämföra. Författarens förhoppning är dels att dessa eller liknande mått kommer att användas av andra forskare inom detta område för att i framtiden kunna utföra djupare och mer omfattande analyser än vad som görs idag. Dessutom kan kanske den genomförda mer detaljerade felanalysen, med traditionella utvärderingsmått, inspirera andra forskare att göra liknande analyser i motsats till de mer ytliga utvärderingar som görs idag. Då skulle en taggares styrka och svagheter framgå mer tydligt.

6 Diskussion om faktorer som påverkar prestation

Detta avsnitt behandlar vissa faktorer som påverkar prestationen vid ordklasstaggning dels generellt och dels med de givna förutsättningar som finns i avhandlingen.

Som angivits tidigare kan påverkande faktorer indelas i två grupper, direkta och indirekta faktorer, dvs. efter sättet de påverkar prestationen.

I avsnitten nedan diskuteras faktorerna mer ingående. Relevanta studier och exempel på utvärderingar av taggningsverktyg som rör de givna faktorerna redovisas, där prestationsmått anges antingen med felprocent eller korrekthet. Felprocent är hur många procent av 100 som inte är korrekta. För vidare beskrivning av dessa mått, se avsnitt 2.2.2.3. Sist följer slutsatser om påverkande faktorer i avsnitt 6.3.

6.1 Direkta faktorer

Av de faktorer som påverkar prestationen i ordklasstaggning på ett direkt sätt diskuteras följande: innehåll och storlek av träningsmaterial, enhetsförråd och informationsstruktur i lexikaliskt material och val av disambigueringsmetod i taggningsverktyg.

6.1.1 Träningsmaterial

Träningsmaterialets innehåll och storlek har stor betydelse för vilka möjligheter som finns för att skapa korrekt regelmaterial som kan användas vid disambiguering. Kvalitet och storlek är de aspekter som normalt diskuteras när det gäller träningsmaterial.

Kvalitet har att göra med hur korrekt och hur fullständigt materialet är. Om det finns felaktigheter i taggningen i materialet följer dessa med i de disambigueringsregler som skapas. Om materialet är ofullständigt skapas färre disambigueringsregler.

I en studie av Gunnel Källgren (1998) undersöktes vilka fel som uppstod i taggningen av en textkorpus. Hon menar att för att komma till rätta med problem som uppstår vid taggning behöver felen analyseras grundligt och inte avfärdas som endast en viss typ av taggningsfel som kan åtgärdas t.ex. genom mer träning. Hon jämför bl. a. taggningsfelsfrekvens mellan manuell och automatisk uppmärkning. I tabell 44 nedan visas resultaten av jämförelsen.

	N	Percent
Errors only in automatic tagging	3591	7,1%
Errors only in manual tagging	503	1,0%
Errors in both	110	0,2%
Total	4202	8,3%

Tabell 44. Fördelning av taggningfel mellan manuell och automatisk uppmärkning, Källgren (1998).

Källgren delade in felen i grupperna 'innehållsord' och 'funktionsord'. Innehållsord är t.ex. nomen, verb, adjektiv, particip och egennamn. I denna grupp uppträder en liten del av alla fel men vanliga fel är nomen som får fel numerusform och adverb som förväxlas med vissa adjektiv. Funktionsorden innehåller den största delen fel. I tabell 45 framgår vilka ord som mest frekvent taggades fel.

Ord	Frekvens
det	330
ett	254
som (rel.pron,adv.)	180
den	153
om	122
en	109
att	83
så	79
ut	73
för	70

Tabell 45. De 10 mest frekvent feltaggade funktionsorden, Källgren (1996).

Orden i tabell 45 kan jämföras med motsvarande siffror från utvärderingen av EPOS i tabell 46 (även tabell 70).

Ord	Frekvens
som	237
om	136
en	133
andra	79
de	79
så	72
att	55
med	52
för	50
den	48

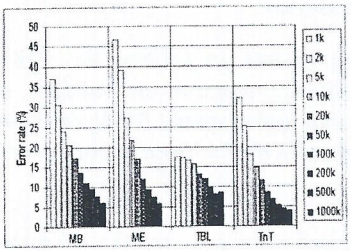
Tabell 46. De 10 mest frekvent feltaggade orden från utvärderingen av EPOS.

Några problem som rör kvalitet och fel i korpusar uppstod också i avhandlingsarbetet. De problem som skapades i och med att SUC användes som träningskorpus för EPOS låg främst i mappningen. Själva SUC-korpusen är i sig en bra korpus att utgå ifrån då träning ska utföras på ordklasstaggare. Men då det i avhandlingen fanns en given lexikalisk resurs (SMDB), som korpusen skulle anpassas till genom mappning, uppstod problem. Problemen var flera. Det rörde sig om konflikter vid olika val av ordklasstaggar för olika ord, och skapande av ny homografi i SUC då det fanns olika utförlig grammatisk representation för ord. Dessutom upptäcktes felaktiga taggningar i SUC. Sammanlagda antalet problem av den ovan nämnda karaktären var ca 6400, ca 0,64% av hela SUC-korpusen. Dessa diskuterades utförligt i avsnitt 4.1.

Storlek eller kvantitet av träningsmaterialet är direkt relaterat till hur stor mängd disambigueringsregler som kan skapas. Ju fler regler, desto bättre taggning, förutsatt att materialet har sådan variation att disambigueringsreglerna både blir täckande och representativa.

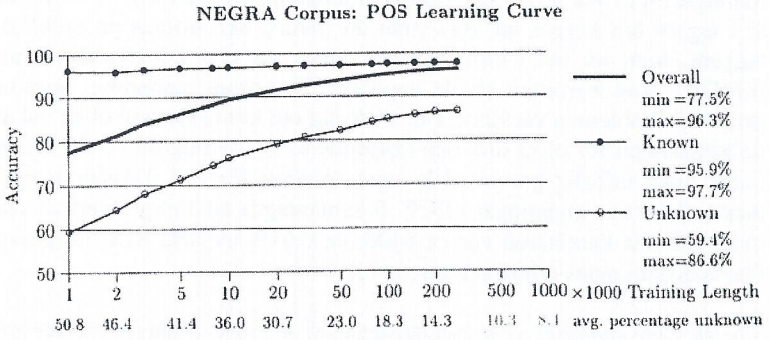
I avhandlingen har inga egna experiment utförts med avseende på storlek av träningskorpus. Det finns andra studier om dess påverkan på prestationen, t. ex. Megyesi (2002) och Brants (1998). I avhandlingen har fördelningen 10% testdata och 90% träningsdata tillämpats på SUC-korpusen.

Megyesi (2002) har i en studie, vilken redogörs för närmare i avsnitt 6.1.3, undersökt vilken effekt olika stora träningskorpusar har på prestationen, se figur 30 nedan, vilken illusterar resultaten av denna studie. Resultaten visade att ju större träningskorpus som användes, desto mer förbättrades prestationen. Maximum Entropy (ME) var den metod som påverkades mest av större träningskorpus, då prestationen förbättrades med 88% vid utvidgning av korpusen från 1000 till 1000000 ord.



Figur 30. Diagram över förhållandet mellan prestation och storlek på träningskorpusar, Megyesi (2002).

Thorsten Brants (1998), visar i manualen till TnT-tagger hur inlärningskurvan kan se ut vid ett exempel med den tyska NEGRA-korpusen (Skut et al. 1997), återgiven i figur 31.



Figur 31. Inlärningskurva vid träning av TnT-tagger på NEGRA-korpusen, Brants (1998).

Brants anger att resultat vid ordklasstaggning idag (1998) ligger mellan 96-97%, beroende på språk och tagguppsättning. Han påpekar också att korrektheten för taggning av okända ord vanligtvis är betydligt lägre än den för kända ord, vilket är ett väntat resultat. Det kan skilja ca 10% i prestation mellan de båda. Han menar att det ofta är lättare att höja korrektheten för taggning på kända ord. Det krävs ibland inte mer än ca 1000 tokens för att uppnå ca 95-96%. I TnT-tagger används vid disambiguering av homografer inte enbart den bästa (mest frekventa) taggen, det finns också en möjlighet att taggaren kan välja en annan tagg vilken är nästan lika frekvent. Brants kallar detta för *säkra* och *osäkra* tilldelningar av taggar. Han har föredragit att returnera en del homografifall i utdatan hellre än att låta taggaren välja en osäker tagg vid disambiguering.

6.1.2 Lexikaliskt material

För ordklasstaggning fordras ett lexikon för skriftspråkets uttryckssida med tillräckligt detaljerade uppgifter om ordklass, böjning och variation. Detta ska användas som utgångspunkt för disambigueringen av homografer. Ett lexikon kan påverka prestationen genom sitt enhetsförråd och informationsstruktur.

Med *enhetsförråd* avses vilka ord som finns representerade i lexikonet. Endast befintliga ord kan disambigueras. Ca 50% av orden i en text behöver disambigueras, övriga ord är heterografer. De ord som inte finns i lexikonet måste genomgå en morfologisk analys och därigenom försvåras uppgiften att ordklasstagga korrekt. Lexikonet påverkar också prestationen

genom sitt utseende.

Med *informationsstruktur* menas vilken morfologisk och syntaktisk information som finns om varje ord. Beroende på vad det är för typ av ord, heterograf eller homograf, kommer resultatet se olika ut. Eftersom resultatet av taggning av heterografer är givet kommer de inte att behandlas här. Svårigheten med homografer, både externa och interna, är att de har minst två taggar för varje ord. Ju fler taggar ett ord har desto större är sannolikheten att ett ord får fel tagg, om man utgår ifrån vilken tagg ett ord slumpmässigt kan få. Men det stämmer inte alltid med faktiska förhållanden, då vissa homografer kan vara ord som har många belägg i en träningstext och därmed också får bra och representativa disambigueringsregler. Generellt sett kan man dock dra slutsatsen att ju fler taggar som existerar per graford i ett lexikon, desto svårare blir disambigueringsuppgiften.

Studier om homografi i ett lexikon är ovanliga, men däremot diskuteras homografi i texten. Karlsson (1992) uppger i arbetet med SWETWOL en homografi på 57,3% i texten medan det i andra källor som t.ex. NFO (1970) anges att homografin ligger högre, 64,5%. Detta menar Karlsson beror på att homografin står i relation till hur ordklassbegreppet definieras. Författarens slutsatser från avhandlingsarbetet visar att homografin i en text dessutom är direkt beroende av hur utförligt ord beskrivs i ett lexikon med avseende på intern och extern homografi.

I brist på annat lexikaliskt jämförelsematerial har SMDB, som lexikalisk resurs, jämförts med ett lexikon skapat från SUC-korpusen (Megyesi, 2002). SMDB har i genomsnitt 1,39 taggar per graford medan SUC-lexikonet i genomsnitt har 1,07 taggar per graford. Det är uppenbart att ett lexikon skapat utifrån en korpus inte har lika många taggar per graford som ett fullständigt lexikon, men trots detta är det ändå vanligt att använda lexikon av denna typ. Detta innebär att det slumpmässigt är 23% större chans att ett ord får fel tagg som taggats med SMDB än om det taggats med SUC-lexikonet.

6.1.3 Disambigueringsmetod

Disambiguering av homografer utförs enligt en viss teknik och metod. Teknik och metod fungerar olika bra för olika språk. Därför är det viktigt att välja en metod och teknik som både praktiskt och språkspecifikt passar den disambigueringsuppgift som ska utföras.

Tekniken[21] kan vara automatisk eller manuell. Idag är det vanligare att tillämpa automatiska tekniker på grund av textkorpusarnas storlek (oftast

mer än en miljon ord). Voutilainen (1999) hänvisar till skillnaden mellan lingvistiska angreppssätt och dataimplementerade angreppssätt. Han menar att lingvistiska motsvarar manuellt skapade taggningsregler och det dataimplementerade är ett resultat av automatiskt utförda statistiska studier av stora textmängder. Machine learning är en teknik som maskinellt automatiskt extraherar disambigueringsregler vilka används som underlag vid disambiguering. I vilket format disambigueringsreglerna sammanställs och lagras är förknippade med vilken metod som sedan ska tillämpas.

Författarens erfarenheter med val av teknik innefattar både manuellt utformade taggningsregler och maskinellt sammanställda regler. Försöket med handskrivna regler redovisades tidigare i avsnitt 4.2.2.2 och avsnitt 4.2.3. Förutom låg prestation var nackdelen att detta angreppssätt krävde lingvistiska ställningstaganden som inte låg inom ramen för denna studie. Däremot antogs att användning av Brill tagger skulle leda till en bättre förutsättning för disambigueringuppgiften. Taggarens prestanda visade sig vara relativt låg, då den kunde uppmätas till ca 93% korrekt disambiguerad text. De flesta normalpresterande taggare idag klarar av ca 95% beroende på utformning av tagguppsättning, disambigueringsmetod, träningskorpus och språk som används.

Metod som används vid disambiguering har att göra med hur det extraherade disambigueringsmaterialet används.

I ett område som ständigt utvecklas, har McEnery och Wilson (1997) har gjort ett försök att utvärdera och sammanfatta några metoder i taggningsansatser som har gjorts i olika språk.

I tabell 47 anges tidstypiska POS-taggare och POS-systemens felprocent.

Skapare	**År**	**Felprocent**
Greene and Rubin	1971	23 %
Garside	1987	4 %
De Rose	1991	4 %
Cutting et. al.	1992	4 %
Brill	1992	5 %

Tabell 47. Prestation i form av procent av fel i POS-taggning, McEnery och Wilson (1997:126).

Ett sätt att utvärdera aktuella tekniker och metoder är att tagga samma text med verktyg som bygger på några av de tidigare nämnda automatiska teknikerna. En studie av Zavrel och Daelemans (1999) har gett följande resultat, se Tabell 48.

21. Här upprepas definitionen av lingvistiska och dataimplementerade angreppssätt för tydlighets skull. Det nämndes tidigare också i kapitel 4.

Taggare	Korrekthet (%)
Rule-based Learner	95,5
Trigram	96,1
Memory-Based	97,0
Maximum Entropy	97,4

Tabell 48. Studie av olika taggares prestation på samma text, Zavrel och Daelemans (1999).

Olika metoder har visat sig passa mer eller mindre bra för olika språk. Zavrel och Daelemans (1999) har gjort en studie på olika språk med olika stora tagguppsättningar men med samma teknik, MBL. Se Tabell 49.

Språk	Tagguppsättningsstorlek	Ant. ord *1000		Korr. (%)
		Träning	Test	
Engelska-WSJ	44	2000	200	96,4
Engelska-LOB	170	931	115	97,0
Holländska	13	611	100	95,7
Tjeckiska	42	495	100	93,6
Spanska	484	711	89	97,8
Svenska	23	1156	11	95,6

Tabell 49. Studie av olika språk men samma taggningsmetod, Zavrel och Daelemans (1999).

Studien i Tabell 49 ger inga indikationer om andra tekniker har så stor spridning mellan olika språk som maximalt 97,8% och minimalt 93,6%, vilket ger en skillnad på 4,2%, men det bevisar ändå att prestationen är språkberoende med avseende på tekniken MBL. Man kan anta att liknande skillnader i resultat kan uppnås även med andra tekniker.

Av alla studier som har gjorts med avseende på prestation i ordklass-taggning gäller de flesta engelska. Beata Megyesi är dock en av dem som har undersökt automatiska taggningsmetoder och deras prestation på svensk text. Det förefaller vid denna tidpunkt vara den mest omfattande studien, både vad det gäller förhållandet mellan prestation och metod, tagguppsättningsstorlek samt storlek på träningskorpus. Megyesi (2001), jämför fyra kända algoritmer med syftet att hitta för- och nackdelar med metoderna, ange vilka fel som skapas, effekterna av tagguppsättnings-storlek samt storlek på träningsmaterialet. De taggare vilka undersöks i studien är MBL, ME, TBL samt TnT (Trigrams'n'Tags-HMM). Samtliga uppges vara tagguppsättningsoberoende och språkoberoende. Tabell 50 nedan visar fördelningen i prestation mellan de olika metoderna, med en

tagguppsättning på 139 taggar och en träningstext på 100 000 token.

Korrekthet	MBL	ME	TBL	TNT
TOTALT %	89,28	91,20	89,06	93,55
KÄNDA %	92,85	93,34	94,35	95,50
OKÄNDA %	68,65	78,85	58,52	82,29

Tabell 50. Prestation hos olika taggare med samma tagguppsättning och träningstext, Megyesi (2001).

Korrekthetsmåttet motsvaras av definitionen av *recall*, se avsnitt 5.1.

TnT bygger på en statistisk trigram-metod, HMM, och visar bäst resultat av de fyra metoderna. Megyesi uppger att metoderna TnT och ME ofta gör taggningsfel som gäller ordklass vid homografa ord medan det med metoderna MBL och TBL ofta görs fel som rör morfologisk kategori, t.ex. numerus. I studien nämns sammanfattningsvis att de resultat som presteras med de olika metoderna inte kommer upp i de nivåer som kan påvisas för engelska. Detta kan enligt Megyesi bero på att svenska är ett morfologiskt rikare språk än exempelvis engelskan.

Egna erfarenheter om metoder i taggningsverktyg är att trots att samma metod har använts i två olika taggningsverktyg (EPOS 1.0 och EPOS 2.0) skiljer sig resultaten i prestationen med ca 2,5 %, vilket gör att man kan dra slutsatsen att andra faktorer har större betydelse i detta specifika fall än taggningsmetod.

Av dessa studier framgår att slutsatser om val av teknik och metod är mer spekulativa då det inte går att jämföra metoder tillämpade på olika språk, Det går inte heller att jämföra metoder med olika grundförutsättningar såsom storlek och innehåll i tagguppsättning och storlek på träningskorpus.

Val av teknik har vid taggning betydelse för vilken täckning taggningsreglerna får. Genom egna erfarenheter blir författarens slutsats att för en enskild individ lönar det sig inte tidsmässigt att implementera taggningsregler manuellt. Det är för tidkrävande med tanke på vilken täckning dessa regler kan uppnå. Det finns givetvis en anledning till att just det datamaskinella angreppssättet är så populärt och används så framgångsrikt i dagens forskning.

Val av metod kan ha betydelse för prestationen hos en taggare. För att komma fram till vilken metod som passar bäst för de givna förutsättningar som finns i ett enskilt fall, t.ex. språk, tagguppsättning och träningskorpus, bör flera metoder prövas och jämföras. Författarens resultat från utveck-

ling av EPOS 1.0 och EPOS 2.0 visar att val av metod inte är helt och hållet avgörande för prestationen. Det finns andra faktorer vilka har större betydelse, t.ex. homografifrekvens i lexikonet. Megyesi visar dock att metoden kan ha stor betydelse i tabell 50, där prestationen skiljer sig åt med 4,49 procentenheter mellan den högst presterande och den lägst presterande metoden.

De tidigare nämnda direkta faktorerna, lexikaliskt och textuellt utgångsmaterial har som också konstaterats haft stor betydelse och påverkat prestationen vid arbetet med EPOS, och eftersom metoden är densamma trots att prestationen skiljer sig åt mellan version 1 och 2, så drar författaren slutsatsen att metoden har mindre betydelse än de andra nämnda faktorerna. Hur de indirekta faktorerna påverkar prestationen diskuteras i följande avsnitt.

6.2 Indirekta faktorer

Av de faktorer som påverkar prestationen i ordklasstaggningen på ett indirekt sätt diskuteras följande: storlek och innehåll i tagguppsättning och utvärderingsmetod. Faktorerna diskuteras i följande avsnitt.

6.2.1 Tagguppsättning

En tagguppsättnings sammansättning kan påverka prestationen på ett indirekt sätt, framförallt vid utvärderingen, då tagguppsättningens storlek och innehåll har betydelse. Språk är unika och har unika morfologiska egenskaper, vilka måste representeras i en tagguppsättning.

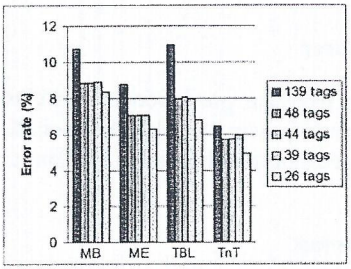
Tagguppsättningens *storlek* diskuteras ofta, eftersom det är en enkel uppgift att räkna antalet taggar. Beroende på vad syftet med taggningen är representeras olika mycket information i tagguppsättningen. Vid utförlig representation blir tagguppsättningen större och vice versa.

I en studie av Megyesi (2001), som gäller betydelse av innehåll och framför allt storlek på tagguppsättning, anger hon att med metoderna MB och TBL görs ofta fel som rör morfologisk kategori, t.ex. numerus. Hon menar att sådana problem kan korrigeras genom att man använder en tagguppsättning med fler morfologiska särdrag än vad som ska utvärderas både vid träning och taggning. Resultaten av ett försök med den metoden visas i Tabell 51.

Korrekthet	MB	ME	TBL	TnT
TOTALT %	92,28	93,49	92,39	95,31
KÄNDA %	94,69	94,72	95,63	96,53
OKÄNDA %	78,37	86,39	73,70	88,24

Tabell 51. Prestation hos olika taggare med större tagguppsättning vid träning och taggning än vid utvärderingen, Megyesi (2001).

Hon anser att denna metod[22] kan tillämpas då ordklasstaggning med färre morfologiska särdrag än vanligt krävs, i tillämpningsområden som t.ex. *informationsutvinning* (*Information Extraction*). I figur 32 redovisas ett diagram från Megyesis studie av olika stora tagguppsättningar och relationen mellan dessa och prestationen med de fyra metoderna.



Figur 32. Diagram vilket illustrerar relation mellan prestation och tagguppsättningsstorlek, Megyesi (2001).

Dan Tufis (2000) undersökte han hur man för språk med många böjningsändelser, t.ex. ungerskan, kan beskära en tagguppsättning på 2148 taggar till 119 taggar och underlätta träning och taggning. Han kallar metoden *tiered tagging*. Genom att steg för steg ta bort vissa attribut vilka beskriver morfologisk information för olika ordklasser kunde han kraftigt reducera tagguppsättningen. De attribut som kunde elimineras var inte avgörande för disambiguering och kunde återskapas vid behov genom information i kontexten. I tabell 52 sammanfattas Tufis resultat av taggning med två olika HMM-taggare, TnT-tagger (Brants 1998) och MULTEXT-ISSCO bigram tagger (Gilberg och Armstrong 1995).

22.Metoden kallas "tiered tagging" och beskrivs nedan efter Tufis (2000).

Tagguppsättning	Tagguppsättningsstorlek	Felprocent (%)	
		MULTEXT	TnT
F (ursprungligt)	2148	6,04	3,92
M (medium)	384	5,65	3,34
O (optimalt)	240	5,45	2,75
R (reducerat)	119	6,00	2,51

Tabell 52. Resultat från Tufis studie med tiered tagging, Tufis (2000).

Tufis konklusion är att man genom att variera storleken på tagguppsättning kan dra viktiga slutsatser oberoende av kontextuell information, genom att studera spridningen av homografi inom olika ordklasser och deras täckning av textord i en korpus.

Thorsten Brants (1995), utförde ett liknande experiment, där han visade att reducering av en tagguppsättnings storlek inte behöver betyda att korrektheten sjunker, snarare tvärtom.

Elworthy (1995), undersökte också storleken på en tagguppsättning och dess påverkan på prestationen för svenska. Elworthys studie skiljer sig från de övriga då han reducerade en tagguppsättning genom att ta bort information om vissa morfologiska egenskaper, t.ex. genus, numerus, bestämdhet och kasus. I tabell 53 nedan visas resultaten från hans studie.

Index	Tagset	Size of tagset	Degree of ambiguity (%)	Ambiguous word accuracy (%)	Unknown word accuracy (%)
1	GNDC	194	52,60	91,09	23,42
2	GnDC	170	50,56	91,62	26,28
3	GNDc	167	52,59	91,01	24,17
4	GNdC	162	52,48	90,77	28,48
5	gNDC	152	52,57	91,19	29,33
6	GnDc	147	50,55	91,51	26,48
7	GndC	141	48,86	91,40	36,29
8	GNdc	140	52,46	90,71	28,63
9	gNDc	134	52,56	91,11	29,48
10	gNdC	126	52,45	90,43	36,14
11	Gndc	123	48,85	91,32	36,44
12	gnDC	121	50,46	91,02	34,73
13	gNdc	113	52,43	90,42	36,24
14	gnDc	105	50,45	90,94	35,39
15	gndC	96	48,75	91,09	48,00
16	gndc	86	48,74	91,02	47,85

Tabell 53. Elworthys studie av relation mellan tagguppsättning och korrekthet, Elworthy (1995).

Versal markerar förekomst av en viss morfologisk egenskap, gemen markerar avsaknad av densamma. Elworthys slutsatser av studien var att korrektheten för okända ord var anmärkningsvärt låg och vissa delar av taggningen borde utföras genom manuell taggning. Han påpekar också att om i de fall en text taggas manuellt bör största möjliga tagguppsättning användas eftersom det vid behov kan reduceras.

En tagguppsättnings *innehåll* är direkt relaterad till den ovan nämnda storleken. Innehållet är beroende av vad syftet med taggningen är. Anledningen till att innehållet oftast inte diskuteras är att det är språkberoende. Ändå borde det vara en självklarhet att vid jämförelse av tagguppsättningar inte bara jämföra antal taggar, utan också innehållet i dem, eftersom det är just innehållet i taggarna som avgör svårigheten att koppla ihop dem med ord vid disambiguering av homografer.

Vid en jämförelse av antal adjektiv i tagguppsättningarna i SUC och i SMDB har både SUC och SMDB, vilket angavs i tabell 19, 37 olika adjektivtaggar. Detta verkar vid första anblicken likvärdigt. Ett konkret exempel som t.ex. ordet 'röda', taggas i SUC med någon av två taggar: 'JJ POS UTR/NEU SIN DEF NOM' eller 'JJ POS UTR/NEU PLU IND/DEF NOM'. I SUC urskiljs varken i singular eller plural om ordet är utrum- eller neutrumform. I SMDB finns följande möjliga taggar för ordet: 'AQPUSND', 'AQPNSND' och 'AQP*PN*'. Även om det finns lika många adjektivtaggar i tagguppsättningen så existerar ändå en högre homografi för vissa adjektiv i SMDB än vad som är möjligt i SUC. Detta är en fråga som har att göra med hur tagguppsättningen används i lexikonet för att representera olika ord. Slutsatsen härav visar att det inte går att tala endast om storlek, utan också om vilka morfologiska särdrag som representeras och hur svåra de är att särskilja sinsemellan.

Slutsatser av studier och arbete med tagguppsättningars storlek och innehåll är att jämförelse mellan tagguppsättningar vanligtvis sker med angivelse av antal taggar. Det bör dock finnas utrymme att föra diskussionen vidare på ett djupare plan där svårigheter med urskiljning av olika morfologiska särdrag förs fram och jämförs, även mellan olika språk.

Megyesi (2002), Tufis (2000) och Elworthy (1995) har visat att det är möjligt att höja eller behålla samma prestation genom att efter träning och disambiguering sedan eliminera viss morfologisk information och göra tagguppsättningen mindre. Denna metod kan tillämpas framgångsrikt beroende på hur specifikt en tagguppsättning behöver vara, vilket i sin tur beror på syftet med taggningen.

6.2.2 Utvärderingsmetod

Eftersom prestation av ordklasstaggning i olika former med olika träningskorpusar, lexikaliska material och disambigueringsmetoder ska mätas är det vanligt att använda ytliga utvärderingsmått, t.ex. korrekthet, recall och precision. Dessa mått kan fungera bra beroende på vad syftet är med att mäta prestationen. Syftet kan t.ex. vara att mäta effekten av förbättringar av disambigueringsmetod med samma tagguppsättning och träningsmaterial som vid tidigare utvärderingar. Det kan också vara att mäta förbättringar eller förändringar som gjorts i tagguppsättning eller träningskorpus. Vid egna studier där endast en viss del av utgångsmaterialet för disambiguering har ändrats fungerar det bra att använda traditionella ytliga prestationsmått.

Men det finns situationer då det är tveksamt om det är möjligt att använda måtten för att göra andra typer av jämförelser. Det kan vara då två olika taggningsmetoder ska jämföras, men då övriga utgångsmaterial inte stämmer överens, t.ex. olika stora tagguppsättningar, olika (stora) träningskorpusar eller olika språk. Då anser författaren att dessa mått inte räcker för att göra en adekvat jämförelse. Det är dock vanligt att göra jämförelser på detta sätt trots att förutsättningarna inte är desamma. Därför är en diskussion nödvändig om hur olika faktorer som påverkar prestation kan representeras och användas i prestationsmått liknande de ytliga som används idag.

I avhandlingen har föreslagits ett sätt att väga in en av faktorerna som kan anses påverka prestation, nämligen homografi. I en fortsättning på avhandlingen kan samma resonemang även användas för att väga in övriga påverkande faktorer såsom träningskorpus, tagguppsättning och disambigueringsmetod. Att väga in homografi som en påverkande faktor i prestation kan naturligtvis göras på olika sätt, och den föreslagna metoden mäter endast homografins utbredning och frekvens och inte andra svårigheter som t.ex. att homografer kan vara olika svåra att disambiguera oavsett hur många taggar det homografa ordet har.

6.3 Slutsatser om påverkande faktorer

Med tanke på de varierande resultat som har visats i de olika studier vilka har redovisats ovan, kan man dra följande slutsatser: det går att jämföra olika tekniker med olika stora tagguppsättningar, tränade på olika stora korpusar med olika språk, men det ger knappast metoderna någon rättvis bedömning. I stället bör jämförelser göras för samma språk med avseende på olika tekniker, olika stora tagguppsättningar eller olika stora tränings-

korpusar. Endast då kan rättvisa bedömningar göras av olika metoder eller andra påverkande faktorer. Av de studier som beskrivits ovan uppfyller fem av dessa det kravet, den av Elworthy (1995), Brants (1998), Zavrel & Daelemans (1999), Tufis (2000) och Megyesi (2001). Detta bör man ha i åtanke när redogörelser för diverse taggningsansatser studeras, t.ex. i kapitel 2.

7 Diskussion om undersökning av förbättringsmöjligheter

Målsättningen för disambigueringen av homografer i Språkbanken har varit att med enklast möjliga medel utföra analyser med stokastiska metoder. Som Jurafsky (2000) konstaterar, är det inte möjligt att med endast stokastiska tekniker uppnå önskat resultat, utan att vissa justeringar, sk. *smoothing*, nödvändigtvis görs. Det kan t. ex. vara i form av manuellt utformade regler eller statistiska regler som inriktar sig på speciella problemområden i disambigueringen. Därför beskrivs ett antal sådana ansatser som har gjorts i syfte att förbättra prestationen i EPOS, och därmed undersöka om och hur mycket det går att förbättra disambigueringen och i så fall till vilket pris. På grund av att det även uppstod problem vid träningen av EPOS kommer undersöks också hur mycket träningskorpusen, dvs. SUC (i mappat format), kan förbättras för att få en så bra utgångspunkt som möjligt inför disambigueringen. Vissa delar av förbättringen av mappningen är resultatet av ett samarbete med Berg, Gellerstam och Cederholm, där de övriga har bidragit med förslag till lingvistiska undantagsregler. Exempel och redovisningar i detta avsnitt bygger huvudsakligen på undersökning av en med EPOS disambiguerad version av Press 65, vilket är ett av tidningsmaterialen i Språkbanken. I avsnittet förslås ett flertal utvidgningar av analyser som inte har varit möjliga att utföra inom ramen för avhandlingsarbetet, men som kan utföras i ett mer långsiktigt arbete med ordklasstaggning i Språkbanken.

7.1 Förbättring av mappat träningsmaterial

Det textmaterial i SUC som kan användas direkt för skapande av disambigueringsregler och suffixregler för morfologisk analys av okända ord är givetvis den del av den mappade versionen av texten som är enkel och problemfri att översätta till SMDB:s tagguppsättning. Det gäller t.ex. textord vilka har taggar i SUC:s tagguppsättning som står i ett 1:1-förhållande med taggar i SMDB:s tagguppsättning. De övriga orden, vilka i SUC har taggar som motsvaras av mer än en tagg i SMDB (ambiguösa mappningar) eller helt saknar motsvarighet (ej motsvarande mappningar) , är de som ställer till problem. Detta problem har tidigare diskuterats ingående i avsnitt 4.1.3.

Som en del av avhandlingsarbetet har detta problem delvis kunnat åtgärdas. I samarbetsgruppen har 'ej motsvarande mappningar' diskuterats ett flertal gånger. Vad som har framkommit är framför allt att vissa fall är enklare att lösa än andra. De kategorier som ställer till problem anges i

tabell 54, i förkortad version från avsnitt 4.1.3.2, där problemfall i mappningsproceduren beskrevs.

F.	Ord	SUC	SMDB	Mappad
a) Två av de mest frekventa problemfallen:				
1199	andra	JJ POS UTR/NEU PLU IND/DEF NOM	MO*SND,V0N0A,V0M0A,PI*P*, DI*P,PI*S*,DI*S	AQP*PN*
1074	alla	DT UTR/NEU PLU IND/ DEF	NCUSNI,PI*P*,D0*P	DF*P
b) Fel i genus, kasus				
1	presi-dentval	NN UTR SIN IND NOM	NCNSNI,NCNPNI - genusfel	NCUSNI
1	meters	NN UTR PLU IND NOM	NCUSGI,NCUPGI - kasusfel	NCUPNI
c) Talspråk, eller felaktiga ordformer				
1	kalla	VB PRT AKT	AQPUSND,AQN-SND,AQP*PN*,V0N0A,V0M0A, NCUSNI	V0ISA
d) Annan tolkning av ord ger andra ordklasser				
33	vänster	JJ POS UTR SIN IND NOM	NCUSNI,AQP*SN0	AQPUSNI
36	faktum	NN NEU SIN DEF NOM	NCNSNI,NCNPNI	NCNSND
53	går	AB	V0IPA,NC0000	R0
e) Felaktigt taggat				
4	litas	NN UTR SIN IND NOM	V0N0P,V0IPS	NCUSNI
3	kasus	NN UTR PLU IND NOM	NCUSNI,NCUSGI,NCN-SNI,NCNSGI,NCNPNI,NCNPGI	NCUPNI
2	var	VB INF AKT	NCNSNI,NCN-PNI,NCUSNI,AQP0000,R0,V0M0A,V0ISA,PIUS0,DIUS	V0N0A

Tabell 54. Exempel på ord som skiljer sig i taggning mellan SUC och SMDB.

I tabell 54 anges först förekommande frekvens av problemfall, 'F.', exempelordet, SUC:s tagg, SMDB:s taggar för det specifika ordet och sist den mappade versionen av den i SUC föreslagna taggen. De fem kategorierna i tabell 54, förklaras en efter en; därefter förs en diskussion om möjliga lösningar.

a) Två av de mest frekventa problemfallen tillhör en kategori där flera fall endast är möjliga att åtgärda genom manuell analys av respektive fall. Detta är ett pågående arbete, huvudsakligen genom förslag till lingvistiska undantagsregler av Berg (och i ett tidigare skede även av Gellerstam) och implementering av författaren, där möjliga motsvarigheter identifieras. Slutsatsen i exemplen ovan är att ordet 'andra', taggat i SUC som ett adjektiv, måste vara de fall då 'andra' står som bestämning till ett nomen, t.ex. "andra människor". Det skulle då i SMDB motsvaras av en artikel, närmare bestämt "DI*P". Då frekventa fall såsom detta har en möjlig motsvarande tagg i SMDB, om än identifierad manuellt, så är det möjligt att

infoga och implementera detta som ett undantag i mappningsmodulen.

b) Fel i genus eller kasus hör till kategorin fel i SUC-korpusen. Genom mappningen och jämförelser mellan mappad tagg och möjliga SMDB-taggar har dessa typer av fel uppdagats. De bör korrigeras direkt i korpusen där de förekommer istället för att behandlas av en speciell undantagsregel. Korrigering är nödvändig i alla fall där faktiska stavfel eller taggningsfel förekommer för att undgå att felen sprider sig till disambigueringsregler som skapas från den mappade texten.

c) Talspråk eller felaktiga ordformer är exempel på ord som inte finns i SMDB med samma ordklasskategorier som i SUC. Eftersom SMDB bygger på SAOL som i huvudsak endast innehåller normerande skriftspråksord och ordformer, kommer de heller inte i framtiden att finnas med där. Därför måste dessa fall utgöra undantag vilka inte kommer att ingå i mappningen på grund av att de inte bör representeras som möjliga former för orden i disambigueringsregler och morfologiska analysregler. En konsekvens av detta val är att vid användning av EPOS i ett senare skede kommer dessa ordformer inte att disambigueras såsom talspråk utan i stället ha fått taggar som t.ex. "kalla/AQPUSND (adj utr sing best), AQNSND (adj neut sing best), AQP*PN* (adj plur), V0N0A (verb inf akt), V0M0A (verb imp akt), NCUSNI (nom utr sing obest)". Vilken tagg som då väljs i disambigueringen beror på kontexten i de fall orden förekommer.

d) "Annan tolkning av ord ger andra ordklasser", är en kategori som speglar skillnader i beskrivning av ord mellan SUC och SMDB. Ett par exempel är "faktum" och "går", där "faktum" i SMDB inte anses ha bestämd form, medan det i SUC har det. Ordet "går" är i detta fall den form som ingår i adverbet "i går", där "går" i SUC taggas som adverb även som självständig form, medan det som ingående form i ett adverb i SMDB anses vara ett nomen i nominativ form. Liknande fall kan åtgärdas i mappningen på grund av de faktiskt rapporteras som fel, men i likhet med kategori a) måste det ske genom manuell analys och genom att de behandlas som undantag med specialregler i mappningen.

e) Felaktigt taggade ord, faller delvis under samma kategori som b), förutom att det rör sig om grövre typer av fel än de i b). De kommer inte att ingå som undantag i mappningsmodulen på grund av att de i stället bör korrigeras i SUC-korpusen av anledningen att felen inte ska spridas till disambigueringsreglerna. Detta måste göras manuellt.

Här exemplifieras hur arbetet med den senaste förbättringen av träningskorpusen utförs. Berg har sammanställt 10 förslag till mappningsöversättningar som kan korrigeras och alla utgör undantag i mappningen

(tabell 55).

F.	Ord	SUC	SMDB	Mappad	Ny tagg
1074	alla	DT UTR/NEU PLU IND/DEF	MO*SND,V0N0A,V0M0A, PI*P*,DI*P,PI*S*,DI*S	DF*P	D0*P
240	båda	JJ POS UTR/NEU PLU IND/ DEF NOM	NCUSNI,V0N0A,V0M0A, PF*P*, DF*P	AQP*PN*	DF*P
53	går	AB	V0IPA,NC0000	R0	NC0000
51	nära	PP	R0,AQPUSNI,AQPNSNI, AQPUSND,AQPNSND, AQP*PN*,V0N0A	S	R0
38	Fast	KN	AQPUSNI,AQPNSNI,RP, R0,CS	CC	CS
31	Båda	JJ POS UTR/NEU PLU IND/ DEF NOM	NCUSNI,V0N0A,V0M0A, PF*P*,DF*P	AQP*PN*	DF*P
29	ene	JJ POS MAS SIN DEF NOM	NCNSNI,PIMS*	AQPM-SND	PIMS*
26	söder	JJ POS UTR SIN IND NOM	NCUSNI,R0	AQPUSNI	R0
20	rätta	NN UTR - - -	AQPUSND,AQPNSND, AQP*PN*,NC0000, V0N0A, V0M0A	NCU***	NC0000
12	höstas	NN - - - -	V0N0P,V0IPS,NCU000	NC0000	NCU000

Tabell 55. De senaste 10 ändringarna i mappningsmodulen.

Korrigering av ett exempel

För att visa vilken effekt korrigering i den mappade texten får, kommer ett exempel med ordet "höstas" att användas. Innan korrigeringar om ordet utfördes kunde disambigueringsregler inte skapas p.g.a. att det i mappningen inte fanns någon överensstämmelse mellan taggarna 'NC0000', mappad från SUC och 'NCU000', från SMDB. I två följande exempel visas skillnaden mellan taggning av 34 fall av "höstas" i korpusmaterialet Press 65[23] i Språkbanken (tabell 56 och tabell 58). Tolv fall hittades i SUC som gäller användning av "höstas" som nomen och inget med någon annan tagg. Dessa 12 fall användes bl. a. för att skapa 16 trigramregler, 12 bigramregler och 1 regel som enbart bygger på ordets frekvens.

Tokennummer	Disambiguerad tagg	Lemma	Lookup-typ	Dis. typ	Token
tnr='11660'	msd='V0IPS'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='12655'	msd='V0IPS'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='31244'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='34566'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='3'	orth='höstas'
tnr='37065'	msd='V0IPS| V0N0P| NCU000'	lemma='hösta.1| höst.1'	atype='1'	dtype='00'	orth='höstas'
tnr='65088'	msd='V0IPS| V0N0P| NCU000'	lemma='hösta.1| höst.1'	atype='1'	dtype='00'	orth='höstas'

23. Korpustexten Press 65 används här som exempel trots att det inte finns någon "guldstandard" till den. På grund av att det här rör sig om så få exempel kan dessa bedömas för hand.

Tokennummer	Disambiguerad tagg	Lemma	Lookup-typ	Dis. typ	Token
tnr='67363'	msd='V0IPS|V0N0P|NCU000'	lemma='hösta.1|höst.1'	atype='1'	dtype='00'	orth='höstas'
tnr='79587'	msd='V0IPS'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='106615'	msd='V0IPS'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='111786'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='113099'	msd='V0IPS'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='123899'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='3'	orth='höstas'
tnr='127656'	msd='V0IPS|V0N0P|NCU000'	lemma='hösta.1|höst.1'	atype='1'	dtype='00'	orth='höstas'
tnr='130997'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='3'	orth='höstas'
tnr='155679'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='3'	orth='höstas'
tnr='155776'	msd='V0IPS'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='163747'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='165785'	msd='V0IPS|V0N0P|NCU000'	lemma='hösta.1|höst.1'	atype='1'	dtype='00'	orth='höstas'
tnr='260954'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='273103'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='280507'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='340012'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='340572'	msd='V0IPS'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='503349'	msd='V0IPS'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='564345'	msd='V0IPS'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='616679'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='644818'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='692102'	msd='V0IPS'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='942115'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='3'	orth='höstas'
tnr='1019047'	msd='V0IPS'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='1040304'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='1053551'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='1082195'	msd='V0IPS'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='1172927'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'

Tabell 56. 'höstas' disambiguerat före korrigering i mappning i EPOS.

Tabellerna visar hur 'höstas' är taggat i 34 enskilda fall i Press 65, där alla ska taggas som 'NCU000'. Som sammanfattning av tabell 56, kan konstateras att 5 fall av 34 var korrekt disambiguerade, 12 fick taggen 'V0IPS', 12 fick taggen 'V0N0P' och 5 fall kunde inte disambigueras alls. För att förklara den information som finns i tabellerna angående 'dtype' (disambigueringstyp) i tabell 57, har det gjorts en sammanställning av dessa utgående ifrån tabell 32.

Dis.typnummer	Namn	Reglernas innehåll vid analys
1	Lexikalisk	Lexikalisk information om ordet som analyseras och lexikalisk information från kontexten.
2	Lexikalisk-kontextuell	Lexikalisk information om ordet som analyseras och information om taggar i kontexten.

Dis.typnummer	Namn	Reglernas innehåll vid analys
3	Kontextuell	Enbart information om möjliga taggar för ordet som analyseras och information om taggar i kontexten.
4	Tokenbaserad	Enbart information om den lexikaliska enheten som analyseras.
5	Suffixbaserad	Suffixinformation om det analyserade ordet och ordets möjliga taggar.
9	Morfologisk analys av okända ord	Suffixinformation om det okända ordet och ev. taggar i kontexten.
10	Kongruens	Morfologisk informatin om vissa ordklasser i nominalfraser.
11	Infinitivartikelanalys	Information om verb i kontexten.
12	Bigram	Kontextuell och lexikalisk information om två intilliggande ord och taggar.
13	Analys av 'som'	Kontextuell och lexikalisk information från kontexten.
00	Ej disambiguerat	Ej möjligt att disambiguera

Tabell 57. Disambigueringstyper med namn och innehåll i EPOS.

Angående kommentarer till tabell 56 är det intressant att notera att i kolumnen som anger disambigueringstyp anges i stort sett uteslutande '3' vilken är den mest generella regeln av alla. Det är en regel som endast använder de ordklasstaggar som finns att tillgå både för det aktuella ordet och för kontexten. Den kan alltså användas då det inte finns några andra regler som gäller ett specifikt ord. Disambigueringstyp '00' betyder att ordet inte kunde disambigueras.

Tokennummer	Disambiguerad tagg	Lemma	Lookuptyp	Dis. typ	Token
tnr='11660'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='12655'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='4'	orth='höstas'
tnr='31244'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='34566'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='37065'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='4'	orth='höstas'
tnr='65088'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='67363'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='79587'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='106615'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='4'	orth='höstas'
tnr='111786'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='113099'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='2'	orth='höstas'
tnr='123899'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='127656'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='130997'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='155679'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='155776'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'

Tokennummer	Disambiguerad tagg	Lemma	Lookuptyp	Dis. typ	Token
tnr='163747'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='165785'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='4'	orth='höstas'
tnr='260954'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='273103'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='2'	orth='höstas'
tnr='280507'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='340012'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='340572'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='4'	orth='höstas'
tnr='503349'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='4'	orth='höstas'
tnr='564345'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='616679'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='2'	orth='höstas'
tnr='644818'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='692102'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='4'	orth='höstas'
tnr='942115'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='2'	orth='höstas'
tnr='1019047'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='2'	orth='höstas'
tnr='1040304'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='1053551'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='12'	orth='höstas'
tnr='1082195'	msd='NCU000'	lemma='höst.1'	atype='1'	dtype='3'	orth='höstas'

Tabell 58. 'Höstas' disambiguerat efter korrigering i mappning.

I tabell 57 anges hur resultatet ser ut efter disambiguering av de 34 fall av "höstas" som finns i Press 65, efter korrigering av mappningsregler för "höstas". Då blir 31 av 34 fall korrekta. Resterande 3 fall taggas som 'V0N0P'. De disambigueringstypsregler som har använts denna gång är av typ '2' vilken är en regel som använder det aktuella ordet med taggar i för- och efterkontext. Regeltyp '3' beskrevs tidigare. Typ '4' är en regel som bygger på enbart ordfrekvens. Regeltyp '12' bygger på ordpar, sk. *bigram*, där två intilliggande ord studeras. Sammanfattningsvis kan de fel som uppstår förklaras som sådana fall då kontexten är så pass ovanlig eller i detta fallet inte stämmer överens med de 12 exempel som kunde användas från SUC. Dessa specifika fall anges för ordningens skull i tabell 59.

Tokennummer	Tagg	Lemma	Lookuptyp	Dis.typ	Token
tnr='79579'	msd='DINS'	lemma='en.2'	atype='1'	dtype='2'	orth='Ett'
tnr='79580'	msd='NCNSNI'	lemma='urval.1'	atype='1'	dtype='3'	orth='urval'
tnr='79581'	msd='NCUPNI'	lemma='dikt.1'	atype='1'	dtype='00'	orth='dikter'
tnr='79582'	msd='R0'	lemma='av.2'	atype='1'	dtype='00'	orth='av'
tnr='79583-79584'	msd='NPUSND'	lemma='-.-'	atype='16'	dtype='00'	orth='Carlos Alvarez'
tnr='79585'	msd='V0ISP'	lemma='publicera.1'	atype='1'	dtype='12'	orth='publicerades'
tnr='79586'	msd='S'	lemma='i.2'	atype='1'	dtype='12'	orth='i'
tnr='79587'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='79588'	msd='S'	lemma='av.1'	atype='1'	dtype='2'	orth='av'
tnr='79589'	msd='NPUSND'	lemma='-.-'	atype='6'	dtype='00'	orth='Kjell'
tnr='79590'	msd='NPNSND'	lemma='-.-'	atype='0'	dtype='9'	orth='A'
tnr='79591'	msd='FE'	lemma='-.-'	atype='10'	dtype='00'	orth='.'

Tabell 59. Första exemplet på fel när kontexten är okänd i EPOS.

Tokennummer	Tagg	Lemma	Lookuptyp	Dis.typ	Token
tnr='564342'	msd='PF*PS'	lemma='-.-'	atype='0'	dtype='9'	orth='De'
tnr='564343'	msd='V0ISP'	lemma='visa.2'	atype='1'	dtype='12'	orth='visades'
tnr='564344'	msd='S'	lemma='i.2'	atype='1'	dtype='12'	orth='i'
tnr='564345'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='564346'	msd='S'	lemma='på.1'	atype='1'	dtype='2'	orth='på'
tnr='564347'	msd='NCNSNI'	lemma='-.-'	atype='0'	dtype='9'	orth='Galerie'
tnr='564348-564349'	msd='NPUSND'	lemma='-.-'	atype='16'	dtype='00'	orth='Leo Castelli'
tnr='564350'	msd='R0'	lemma='i.2'	atype='1'	dtype='3'	orth='i'
tnr='564351-564352'	msd='NPNSND'	lemma='-.-'	atype='6'	dtype='00'	orth='New York'
tnr='564353'	msd='FE'	lemma='-.-'	atype='10'	dtype='00'	orth='.'

Tabell 60. Andra exemplet på fel när kontexten är okänd i EPOS.

Tokennummer	Tagg	Lemma	Lookuptyp	Dis.typ	Token
tnr='1040297'	msd='R0'	lemma='när.1'	atype='1'	dtype='12'	orth='när'
tnr='1040298'	msd='NCNSND'	lemma='mål.1'	atype='1'	dtype='00'	orth='målet'
tnr='1040299'	msd='V0ISP'	lemma='ta.1'	atype='1'	dtype='12'	orth='togs'
tnr='1040300'	msd='R0'	lemma='upp.1'	atype='1'	dtype='00'	orth='upp'
tnr='1040301'	msd='S'	lemma='på.1'	atype='1'	dtype='1'	orth='på'
tnr='1040302'	msd='RP'	lemma='ny.1'	atype='1'	dtype='1'	orth='nytt'
tnr='1040303'	msd='S'	lemma='i.2'	atype='1'	dtype='12'	orth='i'
tnr='1040304'	msd='V0N0P'	lemma='hösta.1'	atype='1'	dtype='3'	orth='höstas'
tnr='1040305'	msd='S'	lemma='inför.1'	atype='1'	dtype='3'	orth='inför'
tnr='1040306'	msd='DIUS'	lemma='en.2'	atype='1'	dtype='12'	orth='en'
tnr='1040307'	msd='AQPUSNI'	lemma='ny.1'	atype='1'	dtype='12'	orth='ny'
tnr='1040308'	msd='NCUSNI'	lemma='jury.1'	atype='1'	dtype='00'	orth='jury'
tnr='1040309'	msd='V0ISA'	lemma='bli.1'	atype='1'	dtype='00'	orth='blev'
tnr='1040310'	msd='NCNSND'	lemma='resultat.1'	atype='1'	dtype='00'	orth='resultatet'
tnr='1040311'	msd='R0'	lemma='frikänna.1'	atype='1'	dtype='3'	orth='frikännande'
tnr='1040312'	msd='FE'	lemma='-.-'	atype='10'	dtype='00'	orth='.'

Tabell 61. Tredje exemplet på fel när kontexten är okänd i EPOS.

Vad som framgår vid närmare analys av exempel 1-3 ovan, är att dessa fall inte alls innehåller något ovanligt ord i kontexten. I förkontexten finns 'i' och i efterkontexten finns 'av', 'på' och 'inför'. Beroende på att de senare inte finns i efterkontexten i de exempel på 'höstas' som finns i SUC används en mer generell regel vid disambiguering och då är risken större att disambigueringen blir fel.

Med utgångspunkten 5% korrekta lösningar uppgick resultatet slutligen till 91,2 % korrekta disambiguerade fall efter korrigering av 'höstas' i mappningsreglerna. Detta visar på en förbättring i prestation för detta enskilda ord på 75,5 %, inklusive de fall som inte kunde disambigueras.

Slutsatser från exemplet 'höstas' är att ord med relativt låg frekvens i en träningskorpus inte kan användas för att generera disambigueringsregler

med hög precision. De kan endast representera vissa specifika kontexter. Men som visats kan trots allt några få regler öka precisionen förvånansvärt mycket.

Av problem som angavs i tabell 54, kommer dessa och liknande fall inte att behandlas vidare inom ramen för avhandlingen utan kommer istället att vidare undersökas av Berg samt läggas in som uppdateringsmoment i EPOS vid lämpliga tillfällen i framtiden.

7.2 Förbättring av disambigueringen

Det är inte möjligt att disambiguera homografer korrekt till 100%, och i utvärdering av disambigueringen, framgår vilka problemområden som finns, men redan i detta avsnitt kan redovisas vilka problem som under arbetets gång har identifierats och vilka ansatser till lösningar som framlagts. Endast försök som genererat relativt säkra korrigeringar har använts. Andra ansatser vilka genererat lika många korrekta som felaktiga förslag, har förkastats.

Tre grupper som utgör problem redovisas och vad som har gjorts för att förbättra disambigueringen av dessa problemgrupper. Grupperna innehåller fel som kan indelas i minst två undergrupper: fel som rör överensstämmelse i fraser, samt frekventa fel som gäller vissa ord. Detta gäller: a) kongruens, b) infinitivmärket 'att' och c) ordet 'som'. Dessutom gjordes ett försök att komplettera trigram med d) 'bigram' där de senare var tänkta att användas då trigram-reglerna var för specifika.

a) *Kongruens*. Det är en stor grupp som innehåller många fel i olika grammatiska kontexter. Denna ansats har inledningsvis koncentrerats till nominalfraser i olika former.

Ett exempel är nominalfrasen 'preliminära ritningar' (adjektiv+nomen);

Före analys ser frasen ut på följande sätt, tabell 62:

Tokennummer	Lookuptyp	SMDB-tagg	Lemma	Token
tnr='963'	atype='1'	msd='AQP*PN*'	lemma='preliminär.1'	orth='preliminära'
tnr='963'	atype='1'	msd='AQPNSND'	lemma='preliminär.1'	orth='preliminära'
tnr='963'	atype='1'	msd='AQPUSND'	lemma='preliminär.1'	orth='preliminära'
tnr='964'	atype='1'	msd='NCUPNI'	lemma='ritning.1'	orth='ritningar'

Tabell 62. Exempel före kongruensanalys i EPOS.

Problemet vid disambiguering är att komma fram till vilken form av de tre varianterna av taggar till ordet 'preliminära', som är den rätta. I ett kongruenstest kan man t.ex. ta fram information om genus, numerus och bestämdhet. Bestämdhet kan vara något vansklig att använda som kongruensinformation eftersom det finns en del undantag, t.ex. 'Värnamo/ NPNSGD nygamla/AQPNSND centrum/NCNSNI' (ett exempel ur SUC). Därför används endast numerus och genus som information i kongruenstest. I det ovan nämnda exemplet räcker det att använda informationen om numerus för att avgöra vilken tagg som är rätt i sammanhanget. Efter analys ser frasen i tabell 62 ut på följande sätt, tabell 63:

Tokennummer	SMDB-tagg	Lemma	Lookup-typ	Dis.typ	Token
tnr='963'	msd='AQP*PN*'	lemma='preliminär.1'	atype='1'	dtype='10'	orth='preliminära'
tnr='964'	msd='NCUPNI'	lemma='ritning.1'	atype='1'	dtype='10'	orth='ritningar'

Tabell 63. Exempel efter kongruensanalys med EPOS.

Vid disambiguering av korpusen Press 65, användes kongruensanalysen i 36 477 fall av 1 272 364 token, vilket utgör 2,87%.

För att förbättra disambigueringen är kongruensanalys den modul som kan förbättra resultaten mest. Säkerheten i analysen är mycket stor eftersom modulen innehåller manuellt utformade regler, där endast helt säkra disambigueringsoperationer utförs. Det ingår ingen statistisk information i analysen.

b) *Infinitivmärket*: '*att*' är ett av de ord som frekvent disambigueras fel. Ordet finns på 8:e plats i en lista av de ord som mest frekvent taggas fel i en utvärdering som gjorts på 10% av SUC (tabell 70). Analysen syftar till att undersöka om 'att' förekommer tillsammans med ett verb i infinitivform. Då kan ordet analyseras med relativt hög sannolikhet som ett infinitivmärke. I ett första steg analyseras endast ord som befinner sig i den direkta närkontexten till ordet 'att'. I en utvidgning av denna analys bör även ord analyseras som förekommer med ett eller flera ords mellanrum från 'att'. Den andra möjliga analysen av 'att', som konjunktion, överlåts åt analyser med trigram. Det vore fel att konsekvent dra slutsatsen att om ordet inte är infinitivmärke, enligt analysen, borde det vara konjuktion. Detta till följd av att infinitivmärkesanalysen inte är fullständig. Om en utvidgning görs kan detta negativa kriterium kanske användas.

Ett exempel som kan analyseras med avseende på infinitivmärke ges i tabell 64.

Tokennummer	Lookup-typ	SMDB-tagg	Lemma	Token
tnr='487'	atype='1'	msd='R0'	lemma='på.2'	orth='på'
tnr='487'	atype='1'	msd='S'	lemma='på.1'	orth='på'
tnr='488'	atype='1'	msd='CS'	lemma='att.1'	orth='att'
tnr='488'	atype='1'	msd='Z'	lemma='att.2'	orth='att'
tnr='489'	atype='1'	msd='V0N0A'	lemma='begripa.1'	orth='begripa'
tnr='490'	atype='1'	msd='CC'	lemma='så.3'	orth='så'
tnr='490'	atype='1'	msd='CS'	lemma='så.3'	orth='så'
tnr='490'	atype='1'	msd='I'	lemma='så.4'	orth='så'
tnr='490'	atype='1'	msd='NCUSNI'	lemma='så.6'	orth='så'
tnr='490'	atype='1'	msd='P0000'	lemma='så.2'	orth='så'
tnr='490'	atype='1'	msd='R0'	lemma='så.1'	orth='så'
tnr='490'	atype='1'	msd='V0M0A'	lemma='så.5'	orth='så'
tnr='490'	atype='1'	msd='V0N0A'	lemma='så.5'	orth='så'
tnr='491'	atype='1'	msd='AQPN-SNI'	lemma='mycken.1'	orth='mycket'
tnr='491'	atype='1'	msd='R0'	lemma='mycket.1'	orth='mycket'

Tabell 64. Exempel före infinitivmärkesanalys med EPOS.

Svårigheten i disambiguering av homografen 'att' ligger i att inte enbart identifiera de verb i infinitivform som följer direkt efter artikeln, utan även hitta de verb som befinner sig längre till höger i kontexten. Efter analys ser frasen ut på följande sätt, tabell 65:

Tokennummer	SMDB-tagg	Lemma	Lookup-tagg	Dis.typ	Token
tnr='487'	msd='S'	lemma='på.1'	atype='1'	dtype='1'	orth='på'
tnr='488'	msd='Z'	lemma='att.2'	atype='1'	dtype='11'	orth='att'
tnr='489'	msd='V0N0A'	lemma='begripa.1'	atype='1'	dtype='00'	orth='begripa'
tnr='490'	msd='R0'	lemma='så.1'	atype='1'	dtype='12'	orth='så'
tnr='491'	msd='RP'	lemma='mycket.1'	atype='1'	dtype='1'	orth='mycket'

Tabell 65. Exempel efter infinitivanalys med EPOS.

Vid disambiguering av Press65, användes infinitivmärkesanalysen i 7988 fall av 1 272 364 vilket utgör 0,63%.

c) Ordet '*som*', tillhör gruppen av ord som frekvent taggas fel. Det hamnar på 1:a plats i en lista av de ord som mest frekvent taggas fel i en utvärdering gjort på 10% av SUC (tabell 70). 'Som' är ett ord som är betydligt svårare att analysera och disambiguera. Själva analysen av ordet bygger på liknande metoder som används generellt vid disambiguering. Som underlag för analys används automatiskt extraherade n-gram, med två ord i vänsterkontexten och två ord i högerkontexten, och en vald ordklass för ordet 'som' i just det sammanhanget. Dessutom används ett sannolikhetsmått som anger relativ frekvens för n-gram-regeln. Formel som används för beräkning av sannolikhet är:

$$P(T_i|W_i, (W_{i-2}W_{i-1},W_{i+1}W_{i+2})) = \frac{|W_i, T_i, (W_{i-2}W_{i-1},W_{i+1}W_{i+2})|}{|W_i, (W_{i-2}W_{i-1}W_{i+1}W_{i+2})|}$$

Sannolikhetsformel för n-gram med fyra ord i kontexten i EPOS. Ekv (32)

Formeln i Ekv (32) utläses så att sannolikheten för att ett ord W_i ska ha taggen T_i, givet orden W_{i-2}, W_{i-1} samt W_{i+1},W_{i+2} i kontexten, är summan av alla förekomster av ordet W_i med taggen T_i och orden W_{i-2}, W_{i-1} samt W_{i+1},W_{i+2} i kontexten, dividerat med summan av alla förekomster av ordet W_i och orden W_{i-2}, W_{i-1} samt W_{i+1},W_{i+2} i kontexten.

I SMDB finns fyra möjliga taggar för ordet 'som', CC, CS, PH**0 och R0. I SUC har ordet endast tre möjliga taggar, CC, PH**0 och R0. Detta implicerar en given utökning av analysen av ordet 'som', nämligen att manuellt skapa regler för att hitta de fall då taggen CS skall användas. I en utvidgning av denna studie åstadkoms detta genom att studera alla 19656 fall med 'som' i SUC. Tills vidare används de automatiskt extraherade reglerna för de övriga tre taggarna.

Ett exempel då analysmodulen för 'som' används ges i, tabell 66:

Tokennummer	Lookup-typ	SMDB-tagg	Lemma	Token
tnr='38'	atype='1'	msd='V0IPA'	lemma='varsna.1'	orth='varsnar'
tnr='39'	atype='1'	msd='DINS'	lemma='någon.1'	orth='något'
tnr='39'	atype='1'	msd='PINS*'	lemma='någon.1'	orth='något'
tnr='40'	atype='1'	msd='CC'	lemma='som.1'	orth='som'
tnr='40'	atype='1'	msd='CS'	lemma='som.1'	orth='som'
tnr='40'	atype='1'	msd='PH**0'	lemma='som.2'	orth='som'
tnr='40'	atype='1'	msd='R0'	lemma='som.3'	orth='som'
tnr='41'	atype='1'	msd='V0IPA'	lemma='likna.1'	orth='liknar'
tnr='42'	atype='1'	msd='DINS'	lemma='en.2'	orth='ett'
tnr='42'	atype='1'	msd='MCNSNI'	lemma='en.1'	orth='ett'
tnr='43'	atype='1'	msd='NCNSNI'	lemma='löfte.1'	orth='löfte'
tnr='44'	atype='10'	msd='FE'	lemma='-.-'	orth='.'

Tabell 66. Exempel före analys av 'som' i EPOS.

Svårigheten i att välja rätt tagg av de tre möjliga för 'som' kan för en människa underlättas genom att man kan byta ut ordet mot t.ex. 'liksom' eller 'såsom' eller 'det som' etc. Det är däremot inte trivialt att utföra samma teknik med automatiska metoder och datorprogram, fast heller inte omöjligt. Det är fullt tänkbart att man skulle kunna hitta synonyma användningar av 'som' genom att på automatisk väg söka i de trigram som redan finns extraherade från SUC-korpusen. Detta skulle i så fall vara en intressant utvidgning av analysen. Efter disambiguering av frasen med 'som' med hjälp av disambigueringstyp '13' (för specifik analys av 'som),

ser exemplet ut som i tabell 67.

Tokennummer	SMDB-tagg	Lemma	Lookuptyp	Dis.typ	Token
tnr='38'	msd='V0IPA'	lemma='varsna.1'	atype='1'	dtype='00'	orth='varsnar'
tnr='39'	msd='PINS*'	lemma='någon.1'	atype='1'	dtype='2'	orth='något'
tnr='40'	msd='PH**0'	lemma='som.2'	atype='1'	dtype='13'	orth='som'
tnr='41'	msd='V0IPA'	lemma='likna.1'	atype='1'	dtype='00'	orth='liknar'
tnr='42'	msd='DINS'	lemma='en.2'	atype='1'	dtype='12'	orth='ett'
tnr='43'	msd='NCNSNI'	lemma='löfte.1'	atype='1'	dtype='00'	orth='löfte'
tnr='44'	msd='FE'	lemma='-.-'	atype='10'	dtype='00'	orth='.'

Tabell 67. Exempel efter analys av 'som' i EPOS.

Vid disambiguering av Press 65 används analysen i 9000 fall av 1 272 364, vilket utgör 0,71%.

d) *Bigram* innehåller endast två ord eller alternativt två taggar. Anledningen till att även denna form av n-gram skapades är att trigram ibland kunde vara för specifika för att det skulle vara möjligt att utföra en analys med ledning av en exakt eller liknande kontext. Jämför fallet som diskuterades i tabell 56-tabell 61. Av tabell 58 framgår att ca. 50% av fallen med 'höstas' kunde lösas med hjälp av bigram. Bigram har disambigueringstypsnummer 12.

$$P\langle T_i|W_i, W_{i+1}\rangle = \frac{|W_i, T_i, W_{i+1}|}{|W_i, W_{i+1}|}$$

Sannolikhetsformel för bigram med ett ord i kontexten i EPOS. Ekv (33)

Formeln i Ekv (33) utläses så att sannolikheten för att ett ord W_i ska ha taggen T_i givet ordet W_{i+1} i kontexten, är summan av alla förekomster av W_i med taggen T_i i kontexten W_{i+1} dividerat med summan av alla förekomster av ordet W_i med ordet W_{i+1} i kontexten.

En exempelfras från Press65 då bigramanalys utförs vid tre olika tillfällen ges i, tabell 68 och tabell 69.

Tokennummer	Lookup-typ	SMDB-tagg	Lemma	Token
tnr='74'	atype='1'	msd='NCUPNI'	lemma='bild.1'	orth='bilder'
tnr='75'	atype='1'	msd='CC'	lemma='som.1'	orth='som'
tnr='75'	atype='1'	msd='CS'	lemma='som.1'	orth='som'
tnr='75'	atype='1'	msd='PH**0'	lemma='som.2'	orth='som'
tnr='75'	atype='1'	msd='R0'	lemma='som.3'	orth='som'
tnr='76'	atype='1'	msd='AF0NSGI'	lemma='strama.1'	orth='stramats'
tnr='76'	atype='1'	msd='V0U0P'	lemma='strama.1'	orth='stramats'
tnr='77'	atype='1'	msd='R0'	lemma='upp.1'	orth='upp'
tnr='78'	atype='1'	msd='NCUSNI'	lemma='med.2'	orth='med'
tnr='78'	atype='1'	msd='R0'	lemma='med.1'	orth='med'
tnr='78'	atype='1'	msd='S'	lemma='med.1'	orth='med'
tnr='79'	atype='1'	msd='DIUS'	lemma='en.2'	orth='en'

Tokennummer	Lookup-typ	SMDB-tagg	Lemma	Token
tnr='79'	atype='1'	msd='MCUSNI'	lemma='en.1'	orth='en'
tnr='79'	atype='1'	msd='NCUSNI'	lemma='en.5'	orth='en'
tnr='79'	atype='1'	msd='PIUS*'	lemma='en.3'	orth='en'
tnr='79'	atype='1'	msd='R0'	lemma='en.4'	orth='en'
tnr='80'	atype='1'	msd='AQPUSGI'	lemma='viss.1'	orth='viss'
tnr='80'	atype='1'	msd='AQPUSNI'	lemma='viss.1'	orth='viss'
tnr='81'	atype='0'			orth='kubisering'

Tabell 68. Exempel före analys med bigram i EPOS.

Bigram har en mindre kontext än trigram och blir därför mer generella. Därmed är det ofta lättare att hitta en motsvarande disambigueringsregel med bigram. I ett experiment undersöktes på vilket sätt prestationen förändrades vid användning av bigram. Det visade sig vara ett bra komplement till mer specifika trigramregler. Bigramreglerna består av två bredvidliggande ord och två taggar och ett probabilistiskt mått.

Tokennummer	SMDB-tagg	Lemma	Lookup-typ	Dis.typ	Token
tnr='74'	msd='NCUPNI'	lemma='bild.1'	atype='1'	dtype='10'	orth='bilder'
tnr='75'	msd='PH**0'	lemma='som.2'	atype='1'	dtype='2'	orth='som'
tnr='76'	msd='V0U0P'	lemma='strama.1'	atype='1'	dtype='12'	orth='stramats'
tnr='77'	msd='R0'	lemma='upp.1'	atype='1'	dtype='00'	orth='upp'
tnr='78'	msd='S'	lemma='med.1'	atype='1'	dtype='1'	orth='med'
tnr='79'	msd='DIUS'	lemma='en.2'	atype='1'	dtype='12'	orth='en'
tnr='80'	msd='AQPUSNI'	lemma='viss.1'	atype='1'	dtype='12'	orth='viss'
tnr='81'	msd='NCUSNI'	lemma='-.-'	atype='0'	dtype='9'	orth='kubisering'

Tabell 69. Exempel efter analys med bigram i EPOS.

Vid disambiguering av Press 65 används analysen i 242 400 fall av 1 272 364, vilket utgör 19,05%. Vid en utvärdering visade det sig att komplettering med bigram-analys höjde precisionen med 1,4%.

Summering

I tabell 70 visas de 10 mest frekvent feltaggade teckensekvenser (både textord och interpunktionstecken), som har framkommit i utvärderingen av SUC. Informationen kan användas för att bestämma var det är mest lämpligt att göra ansatser till förbättringar i disambigueringen.

Ord	Antal feltaggade förekomster
som	237
om	136
en	133
andra	79
de	79
så	72
att	55
med	52
för	50
den	48

Tabell 70. Tabell med de 10 mest frekvent feltaggade orden i EPOS.

Förutom 'som' och 'att' där ansatser till lösningar redan gjorts, kan en mer allmän slutsats dras om de ord som finns med i tabell 70. De ord som är svårast att disambiguera med enbart statistiska metoder är homografa ord vilka används ungefär lika frekvent i minst två olika former. Exemplet 'de' är sålunda mycket frekvent både som pronomen och bestämd artikel. Andra slutsatser som rör utvärderingen behandlas i kapitel 5.

8 Slutsatser

Problemställningen för avhandlingen som presenterades i inledningskapitlet var att undersöka till vilken grad det går att korrekt morfosyntaktiskt uppmärka texter med givna resurser.

För att behandla denna problemställning undersöktes olika faktorer som påverkar prestationen och möjligheter till förbättring av resultatet. Dessutom var det av intresse att veta vilka typer av förbättringar som var mest lönsamma.

Slutsatserna i detta avsnitt behandlas i samma ordning som faktorerna som diskuterats tidigare i kapitel 1 och 6, dvs. indelade i direkta och indirekta faktorer.

8.1 Slutsatser om direkta faktorer

Lexikaliskt och textuellt utgångsmaterial och val av disambigueringsmetod är några av de faktorer som har en direkt påverkan på prestationen. Slutsatser om faktorers påverkan på prestationen redovisas i avsnitt 8.1.1-8.1.3 nedan.

8.1.1 Textuellt utgångsmaterial

De egenskaper som brukar diskuteras ifråga om textmaterial är kvalitet och storlek. Eftersom storlek är enkel att mäta och inte beroende av språk eller metod, bortseende från helt avvikande skriftspråk, har flera studier om detta genomförts.

Författaren har i denna avhandling undersökt dels det textmaterial som har funnits att tillgå vid träning av EPOS, dvs. SUC, och dels textmaterial som taggats med EPOS. Det senare textmaterialet har undersökts genom kartläggning av vilka fel som har uppstått och varför. Detta beskrivs vidare under avsnittet om utvärderingsmetodens betydelse för prestationen, 8.2.2. Källgren (1996) har gjort en liknande studie där hon kritiserade utvärderingar för att vara för ytliga vilket leder forskarna till att försöka förbättra prestation genom mer träning i stället för att ta reda på orsaken till felen.

I kapitel 4 i denna avhandling redogjordes för problem som uppstod med det textuella utgångsmaterialet i samband med mappning av SUC:s tagguppsättning till den tagguppsättning som används i SMDB. Detta är en typ

av problem som sällan diskuteras. Problemen berodde delvis på att det för flertalet taggar inte fanns något 1:1-förhållande mellan SUC:s taggar och SMDB:s taggar. I vissa fall motsvarades en tagg i SUC av två eller flera taggar i SMDB. Därmed uppstod en oväntad homografi som i många fall måste lösas manuellt. Ett annat problem var att vissa ord inte hade samma tagg i SUC som i SMDB. Vid konstruktionen av tagguppsättningar hade man valt olika lösningar. Då kunde inte heller de reguljära mappningsreglerna användas. I de flesta fall kunde dock lingvistiska undantagsregler skapas, vilka gjorde det möjligt att översätta dessa ord med taggar från SMDB:s tagguppsättning trots att de hade andra taggar i SUC (pågående arbete vid institutionen, Sture Berg).

Möjligheter till förbättring av kvaliteten hos det textuella utgångsmaterialet undersöktes och diskuterades i kapitel 7. Där visades att förbättring av den lexikaliska beskrivningen och av de regler som hanterar ett enskilt ord som utgör ett problemfall i mappningen, kunde leda till 75% förbättring av prestationen vid disambiguering av detta enskilda ord.

I SUC 2.0 finns fortfarande en del tryckfel/taggningsfel som inte har korrigerats från den tidigare versionen 1.0, men minst 95% av alla fel har dock korrigerats. En del taggningsfel har upptäckts i arbetet med mappning från SUC:s tagguppsättning till SMDB:s tagguppsättning. Eftersom SUC i ett första steg genomgick en automatisk taggning skapades en del taggningsfel på det sättet. Alla fel upptäcktes inte vid den efterföljande manuella korrigeringen. Dessutom kan det på vissa ställen i materialet märkas en inkonsekvens i taggningen till följd av att den har utförts av ett flertal personer med varierande bakgrund. Inkonsekvens i taggningen kan leda till sämre täckning för disambigueringsregler som automatiskt extraheras från sådant textmaterial.

I kapitel 6 ges exempel på studier av Megyesi (2001) och Brants (1998) som har undersökt träningskorpusens storlek i förhållande till prestationen. De kunde visa på ett klart samband mellan stora träningskorpusar och hög prestation. Eftersom det fanns flera liknande undersökningar både för svenska och andra språk undersöktes inte liknande samband i avhandlingen. På grund av att det endast fanns en korpus att tillgå (SUC), kunde i detta läge inga experiment utföras när det gäller att öka storleken på korpusen.

Homografin i texten har också stor betydelse för prestationen. Karlsson (1992) och Allén (1970) anger 57,3% respektive 64,5% homografa ord i en text. Denna skillnad beror enligt Karlsson på hur ordklassbegreppet definieras. Författarens slutsatser är att homografin i en text visserligen står i relation till hur definitionen av ordklasser har gjorts men den har egent-

ligen sitt ursprung i hur motsvarande lexikon utformas och vilken ordklassmodell, som ligger till grund för utformningen. Homografin är förmodligen också genreberoende. Detta område är såvitt man kan konstatera av litteraturreferenser ännu lite utforskat. Beskrivning av homografi i lexikon görs i avsnitt 8.1.2.

8.1.2 Lexikaliskt utgångsmaterial

Det lexikaliska utgångsmaterialet kan ha en påverkan på prestationen genom sitt enhetsförråd och sin informationsstruktur.

Enhetsförrådet i ett lexikon beskriver vilka ord som finns representerade. Ett stort lexikon ger därför bättre resultat än ett mindre. SMDB har ca 1,2 miljoner former representerade och har därmed en mycket bra täckning. Då SMDB används för uppmärkning görs dessutom en sammansättningsanalys på okända ord, Berg et al. (2003). Övriga okända ord analyseras i en morfologisk analys i EPOS.

Informationsstrukturen i lexikonet skiljer på heterografer och homografer. Det har stor betydelse vilken homograffrekvens ett lexikon har. Detta har i sin tur stor påverkan på prestationen eftersom ju fler homografer ett ord har, desto svårare blir uppgiften att slumpmässigt välja rätt tagg för ett ord. De siffror som jämförs i avhandlingen i kapitel 6 är 1,07 taggar per ord i det av Megyesi genererade lexikonet från SUC och 1,39 taggar per ord i SMDB. Slutsatsen man kan dra av dessa siffror är att det rimligtvis är slumpmässigt svårare att tagga en text korrekt med SMDB som lexikon än med SUC-lexikonet.

I avhandlingen, kapitel 5 har också föreslagits ett nytt sätt att jämföra reguljär prestation i förhållande till hur den påverkas av homografi i ett lexikon. Effekten av att använda detta nya utvärderingssamband, $H_{measure}$, blir att prestationen uppvärderas ju större homografin i lexikonet eller homografin i en text är.

8.1.3 Metod

Den metod som använts i avhandlingen vid konstruktion av EPOS är en relativt enkel statistisk metod som bygger på trigram. Den statistiska analysen har kompletterats med punktinsatser på vissa problemområden. En morfologisk analys har också använts för analys av okända ord. Metoden har i de två versioner av EPOS som konstruerats inte visat sig vara avgörande för prestationen, men det är däremot inte uteslutet att en annan metod kan ge bättre prestation.

De problem som uppstår vid disambiguering är främst relaterade till felaktigt taggade ord i textmaterialet. Fel och problem i taggningen diskuterades tidigare i kapitel 4. Fel kommer även att diskuteras i slutsatser om utvärderingsmetod. De fel som diskuterades tidigare var t.ex. hur talspråk skulle tolkas. SMDB är ett lexikon som är skapat för att endast beskriva skriftspråk. Därför behövs en separat kontroll över ord som eventuellt kan vara talspråk, då dessa annars riskeras att taggas fel. Ett annat problem var de okända orden, vilka ger en relativt hög felprocent. En förbättring av den modul som hanterar okända ord är därför nödvändig. Den bör inte endast bygga på slutsatser som dras av morfologisk ändelseinformation utan också om andra kontextuella faktorer. De vanligaste okända orden är egennamn, utländska ord eller nomen.

Förbättring av prestationen med den metod som använts har undersökts i kapitel 7. Där redovisades vanliga fel och möjliga åtgärder för att korrigera dessa. Punktinsatser som har använts för att korrigera vanliga fel är bl. a. kongruensanalys, analys av infinitivmärket 'att', analys av ordet 'som' och en kompletterande statistisk analys med hjälp av bigram. Av punktinsatserna var den ansats som använde sig av regeltypen bigram den mest framgångsrika. Den genererade endast 3% fel, vilket kan jämföras med vissa generella kontextuella analyser som utfördes i EPOS där mycket lite eller ingen information alls finns att tillgå, då felaktigt genererade lösningar kunde stiga till 39%. Detta redovisas i utvärderingen i kapitel 5.

Forskning inom ämnet ordklasstaggning har visat att förbättring med stor effekt endast kan utföras till en viss gräns. Därefter blir effekten mindre. Slutsatserna av olika försök med förbättringar i avhandlingen är att punktinsatser ändå ger stor effekt på förbättring av prestationen och en utökning av t.ex. kongruenskontroll kommer att vara den mest lönsamma satsningen för att förbättra prestationen.

8.2 Slutsatser om indirekta faktorer

Tagguppsättning och val av utvärderingsmetod är några av de faktorer som har en indirekt påverkan på den uppmätta prestationen.

8.2.1 Tagguppsättning

En tagguppsättning kan påverka en ordklasstaggares prestanda indirekt genom sin storlek och sitt innehåll.

I denna avhandling har inga undersökningar gjorts med avseende på storleken av en tagguppsättning. Däremot har betydelsen av hur graford repre-

senteras diskuterats. Vid jämförelser av storlek och innehåll i SUC:s tagguppsättning och SMDB:s tagguppsättning finns det en uppenbar skillnad i storlek, 153 taggar respektive 133 taggar. Vid närmare undersökningar har författaren kommit fram till att SMDB trots detta i många fall representerar en mer specifik taggning för framför allt adjektiv. Därför kan det i SMDB:s tagguppsättning finnas fler taggar inom adjektiv än det gör i SUC:s tagguppsättning. Därav uppstod en oväntad homografi i mappningen mellan tagguppsättningarna, då en tagg i SUC:s tagguppsättning kunde motsvaras av mer än en tagg i SMDB:s tagguppsättning. I och med att SMDB:s taggning då blir mer specifik är detta ett bevis för att storleken inte är avgörande för hur bra eller hur specifik taggning som kan åstadkommas med en tagguppsättning.

Andra studier av bl. a. Megyesi, Tufis, Brants och Elworthy vilka undersöker hur en tagguppsättnings storlek påverkar prestationen redovisas i kapitel 6. Den naturliga, genomgående slutsatsen av dessa studier är att prestationen stiger när en tagguppsättning minskas i storlek.

Då det är mindre viktigt att representera så mycket morfologisk information som möjligt i en text kan metoden att minska tagguppsättning vid utvärdering tillämpas framgångsrikt. Men eftersom målet med taggning med EPOS är att beskriva så mycket syntaktisk och morfologisk information som möjligt, för att möjliggöra vidare syntaktisk analys, är detta inte ett alternativ.

8.2.2 Utvärderingsmetod

Syftet med en utvärdering är att beskriva prestationen av ett arbete, i detta fall automatisk ordklasstaggning.

Val av utvärderingsmetod är viktigt för att på ett korrekt och rättvisande sätt redovisa prestationen. Ofta används ytliga prestationsmått som t.ex. korrekthet, recall och precision. Dessa mått redovisar inte i djupare detalj utan ger en översiktlig prestationsredovisning, för det mesta med ett gemensamt mått för alla ordklasser.

I avhandlingen har en mer detaljerad redovisning gjorts dels vad det gäller prestation på kända och okända ord, men också dess prestation per ordklass. Bl. a. Elworthy har gjort liknande studier med avseende på kända och okända ord.

En utvärdering måste också utföras och betraktas i förhållande till syftet. Om syftet med taggningen är att lemmatisera, identifiera nyord eller

grammatiskt parsa den taggade texten, så bör också utvärderingen spegla relevant information för syftet. Utvärderingen behöver inte täcka mer än vad som krävs för ändamålet eftersom resultatet av prestationen ofta framstår som sämre ju mer information som ska utvärderas.

I avhandlingen har en kartläggning av vilka fel som uppstår vid taggning utförts, där fel och feltyper har redovisats framför allt för att öka möjligheterna att förbättra prestationen hos taggaren EPOS. Även andra indelningar av fel kan göras, t.ex. genom gruppering av morfologiska fel och kategorifel, jfr. Källgren (1992), vilket redovisas i forskningsöversikten.

I forskningsöversikten och i kapitel 5 ges också en översikt över vanliga mått som används för att bedöma prestation hos en taggare. Eftersom homografi i lexikonet har visat sig ha stor betydelse för prestationen, har ett förslag till ett mått föreslagits som bättre värderar prestationen ju högre homografi det finns i ett lexikon, eller en text. Det finns även ett annat mått, Kappa-måttet, vilket i princip kunde användas för att beskriva svårigheten att slumpmässigt disambiguera homografer med olika många taggar, men detta mått är svårt att relatera till normala prestationsmått.

Författarens slutsatser av utvärderingen är att för att det skall vara möjligt att dra några egentliga slutsatser om feltaggningar och hur en taggare ska kunna förbättras efter några givna förutsättningar, måste en grundlig kartläggning av fel utföras. För att hitta bättre lösningar är det också viktigt att undersöka vilka typer av disambigueringsregler som inte presterar bra, och i vilka situationer de är sämre. När olika förutsättningar råder är det också viktigt att så många påverkande faktorer som möjligt kan mätas. Därför bör utvärderingsmått, liksom det som föreslagits för att mäta homografi, även användas för att mäta storlek på träningskorpus, storlek och innehåll i tagguppsättning och även andra faktorer som påverkar eller försvårar prestationen av taggningen.

8.3 Vidare forskning

Om det arbete som utförts i avhandlingsprojektet, särskilt utvecklingen av ett nytt taggningsverktyg, skulle utföras idag hade författaren sannolikt valt ett annat tillvägagångssätt. Men fruktbara insikter och information har framkommit genom de undersökningar som har gjorts inom avhandlingen och därför står det nu klart vad som behöver göras för att förbättra prestationen av EPOS. För att kunna dra några egentliga och djupare slutsatser om prestation är det alltså nödvändigt att göra en kartläggning av vilka taggningsfel som uppstår och varför.

I vidare forskning inom problemställningens område skulle det vara intressant att undersöka vilka andra statistiska metoder som kan förbättra EPOS prestation. En liknande metodbaserad undersökning har utförts av Megyesi (2002). Det är också viktigt att arbeta vidare på de punktinsatser som föreslagits och utöka täckningen av dessa. För att förbättra analysen av okända ord, kan man t. ex. utöka namnlexikon och listor med utländska ord. För att förbättra analysen av talspråk, kan talspråkslexikon användas, t.ex. genom att utnyttja taggning av en parallellkorpus som innehåller talspråk och länka talspråksvarianter med normaliserat språk.

En annan viktig lexikalisk resurs som är nödvändig framför allt i utvärderingsarbetet, är en förteckning över vanliga flerordsenheter och dess särskrivna motsvarigheter. Det är annars omöjligt att jämföra de två skrivsätten.

För att kunna utföra en djupare och mer detaljerad analys vid utvärdering bör fler utvärderingsmått utvecklas som tar hänsyn till olika faktorer som försvårar ordklasstaggning, t.ex. innehåll och storlek av tagguppsättning, storlek och innehåll i träningskorpus etc.

Homografi i lexikon och text har undersökts och diskuterats i flera omgångar i avhandlingen. Eftersom det inte i någon större utsträckning har undersökts hur homografi i en text är genreberoende för svenska språket, är detta ett område som bör undersökas vidare. Dessutom vore det intressant att undersöka i vilken mån frekvensinformation om mindre vanliga homografer från t.ex. NFO skulle förbättra disambigueringen av homografer. I avhandlingen har denna metod undersökts för de 100 vanligaste homograferna utan större framgång.

När ordklasstaggningen med EPOS har en acceptabel kvalitet planerar författaren att vidareutveckla det redan påbörjade arbetet med att anpassa Steven Abneys Scol-parser så att den kan användas för grammatisk parsning av ordklasstaggade texter med den tagguppsättning som används i EPOS och SMDB.

8.4 Sammanfattning

Avhandlingsarbetet visar att det är möjligt att med relativt enkla statistiska metoder uppnå hög precision.

Ett antal faktorer påverkar dock resultatet. Textuellt utgångsmaterial är en av dessa. Eftersom utgångsmaterialet är ofullständigt föreslås i avhandlingen vissa förbättringar och effekten av dessa bevisas. Hög homografi i

lexikaliskt utgångsmaterial har visat sig ha störst betydelse för skillnaden i prestation mellan EPOS v. 2.0 och dess föregångare v. 1.0. Då den statistiska metoden som används i de båda versionerna av EPOS är densamma dras slutsatsen att metoden i det här fallet inte har avgörande betydelse för skillnaden i prestationen mellan v. 1.0 och v. 2.0. Möjligtvis kan andra statistiska metoder generellt leda till en förbättring av resultatet. Betydelse av storlek och fördelning av tagguppsättning bör inte ha någon avgörande betydelse i skillnaden i prestation mellan de båda versionerna av EPOS på grund av att tagguppsättningen i v. 1.0 skapades för att ett 1:1-förhållande mellan tagguppsättningarna skulle kunna användas. Övriga taggar var expanderade taggar (mer specifika) som egentligen inte fick någon genomslagskraft eftersom de knappt användes i utgångsmaterialet. Det fanns även speciella morfologiska distinktioner inlagda i lexikonet i v. 1.0 som markerade kategorier av adverb och verb, vilket heller inte användes i träningskorpusen.

Vid utvärdering av resultat av taggning har valet av utvärderingsmetod som används avgörande betydelse. Metoden bör för att vara rättvisande ta hänsyn till faktorer som har avgörande betydelse för prestationen. Detta är normalt inte fallet eftersom de vanligaste bedömningsmåtten som används vid ordklasstaggning är mycket ytliga. Faktorer som har avgörande betydelse för bedömning av prestation är storlek och fördelning av tagguppsättning, kvalitet och storlek på träningsmaterial och homografifrekvens i lexikon och text. Ett förslag för att förbättra möjligheten att bedöma prestationen i förhållande till homografi i lexikon och text har gjorts genom en formulering av ett utvärderingsmått för detta ändamål. I en utvidgning till denna diskussion kan de andra faktorerna också ingå i detta mått.

Förbättringar av resultat kan utföras på olika plan. Dels vad gäller textuellt utgångsmaterial och dels i den slutliga disambiguerade texten. Förbättringar i utgångsmaterialet har diskuterats och effekten av sådana har också bevisats. Arbete pågår med ytterligare förbättringar. I slutresultatet har framkommit att punktinsatser på problemområden kan höja precisionen betydligt. Grunden för denna slutsats är bl. a. att en detaljerad felanalys har utförts och därför kan existerande problemområden anses vara kartlagda.

Det finns inget automatiskt taggningssystem som taggar texter till 100% korrekt. Som visats i avhandlingsarbetet finns flera områden där förbättringar kan utföras. Förbättringar är dock tidskrävande och därför kommer arbetet som har inletts i avhandlingen att fortsätta.

9 Summary

In Natural Language Processing (NLP) and different kinds of text processing, Part-of-Speech (POS) tagging is a natural first step. POS tagging can be performed manually, or more commonly, automatically. Approximately 50 per cent of all words in a Swedish text are homographs. This means that the words may have at least two possible POS tags. Depending on how detailed the POS tagging is, there will be a varying degree of *homography* in a text. Sometimes homography within a POS tag will emerge if the morphological information specified is detailed enough.

Homography can be visualized by using a lexicon with more than one POS tag assigned to a word when annotating words in a text. The next step in POS tagging is the disambiguation process. This involves deciding in what context a certain POS tag is correct. Even if homography always exists potentially in a text, it is also dependent on the lexicon used. A small and limited lexicon which does not contain much homography makes the disambiguation process easier. Correspondingly, disambiguation is made more difficult when a larger and extended lexicon is used. From a performance point of view, a limited lexicon is preferred, but from a correctness and completeness point of view, an extended lexicon is preferred.

Apart from the fact that homography in a lexicon and a text makes POS tagging more difficult, correct tagging also depends on the *computational method* used, assuming that it is performed automatically. There are several methods, the most successful ones being based on statistical techniques.

The above-mentioned factors - i.e. homography in the lexicon and the text, as well as computational method have a direct impact on the result of POS tagging, since changes in the definition of homography and in technical methods imply immediate effects. There are, however, other factors which, in a more indirect way, affect how the result is perceived. One such indirect factor is how the POS tags are defined and composed in the tagset used. Another factor is the evaluation method employed.

The purpose of a *tagset* is to describe POS tags. The more specific the morphological information provided in the tags is, the more homography is formed in the lexicon and in the text, making the task of disambiguation more difficult. Depending on the language in which the text is written, the amount of morphological information which needs to be described in a tagset will vary. For example, compare English, in which few inflectional

forms need to be described, with Hungarian, which consists of a large number of inflectional forms. Obviously a smaller and more limited tagset can be used for English, whereas a more complex and extended tagset is needed for Hungarian.

The *evaluation method* used also affects the result of POS tagging indirectly. This is done indirectly since it involves how the result is perceived rather than the result in itself. There are different ways of measuring the correctness of a POS tagged text. Some measures are more general and some are more specific. The general measures provide an overall picture of the performance and are the ones most commonly used. They do not take into consideration various relevant factors. Therefore, more specific methods should be developed which will focus on such factors as influence the level of difficulty involved in disambiguation.

9.1 Purpose

This study aims at examining factors and problems which bear on the accuracy of automatic POS tagging with real examples, taken from work with tagging texts in Språkbanken (the Language Bank of Swedish), with the following three given resources: Svensk Morfologisk Databas (SMDB) (Swedish Morphological Database), En Probabilistisk Ordklasstaggare för Svenska (EPOS) (A Probabilistic POS Tagger for Swedish) and the Stockholm-Umeå Corpus (SUC).

More specifically, the aim of the study is to decide to what degree it is possible to achieve a correct POS tagging with the given resources. This question will be answered by examining various factors which, directly or indirectly, may affect performance. These are the initial lexical and textual materials, method of disambiguation, tagset size and composition, and finally the evaluation method employed. Besides examining these factors there will also be a discussion as to what will be the most effective way to improve performance. Three possible ways are examined: improvement of training materials, statistical disambiguation methods and attempts to improve certain minor problem areas. Finally, there will be a discussion of how performance of correct tagging should be evaluated. This concerns what measures are normally used and how they reflect relevant factors. Possible factors are size and composition of tagsets, quality and size of training materials, and degree of homography in the lexicon and text analysed.

9.2 Conclusions

The conclusions drawn from examining relevant factors which may have an impact on the performance of POS tagging are presented below, where direct and indirect factors are accounted for separately.

9.2.1 Direct factors

Initial lexical and textual materials, as well as choice of disambiguation method, are some of the factors which have a direct effect on performance. Conclusions regarding their influence are summarized in sections 9.2.1.1-9.2.1.3.

9.2.1.1 Initial textual material

The properties normally discussed in connection with text materials are quality and size. Since size is easy to measure as well as language independent, except for some completely deviating written languages, this has been the topic of several studies.

In this work I have examined both the textual material used for the training of the tool EPOS, SUC, and the textual material tagged with EPOS. The latter material has been examined by making a survey of the errors which ocurred, and why. This is further described in the section concerning the importance of the method of evaluation used for assessing performance, 8.2.2. In a similar study, Källgren (1996) criticized evaluations for being too superficial. This will normally lead researchers to try to improve performance by means of more training, instead of trying to find out the underlying reason for the errors.

In chapter 4 the problems which arose in mapping the textual material from SUC's tagset onto SMDB's tagset are reported. These kinds of problems are seldom discussed. They were partly due to the fact that for most of the POS tags there was no 1:1 agreement between SUC's tags and SMDB's tags. In some cases, one tag in SUC corresponded to two or more tags in SMDB's tagset, giving rise to unexpected homography which in many cases had to be disambiguated manually. Another problem was that some words did not have the same POS tag in SUC as it had in SMDB. In composing tagsets, various solutions have been chosen. Thus, it was impossible to use the standard mapping rules to translate one tag into another. In most cases, linguistic exception rules could be constructed, which made it possible to translate these words with tags from SMDB's tagset, although they were represented by other tags in SUC (work in

progress at the department of Swedish Language, Sture Berg).

Possibilities for improving the quality of the textual materials are examined and discussed in chapter 7. A possible improvement of the lexical description, and of the linguistic exception rules which handle a specific word constituting a problem in the mapping, could lead to as much as a 75 per cent improvement of the disambiguation performance for a specific word.

In SUC 2.0 there still remain some misprints and tagging errors which have not been corrected from the earlier version 1.0, but approximately 95 per cent of all errors have been corrected. Some tagging errors have been discovered in the mapping process from SUC's tagset to SMDB's tagset. Since, at an earlier stage, SUC was automatically tagged, some tagging errors resulted from that process. All errors were not detected in the following manual correction. Inconsistency in tagging may also be noticed, due to the fact that there were several people with various backgrounds, involved in the manual correction. Inconsistency in the tagging may lead to lower coverage for the disambiguation rules created from such textual material.

Chapter 6 privides some examples of studies by Megyesi (2001) and Brants (1998), examining the size of the training corpus in relation the performance. They could both establish a clear connection between large training corpora and high performance. Since there are several similar studies for Swedish as well as other languages this connection was not considered of interest to stydy again. Another relevant factor is the fact that there was only one available corpus (SUC) for training and testing, so this time it was not possible to investigate the relationship between size and performance.

Homography in the text is also of great importance for performance. Karlsson (1992) and Allén (1970) give 57.3 per cent and 64.5 per cent, respectively, for the presence of homographs in a text. This difference is, accoding to Karlsson, due to different definitions of the notion "part of speech". My conclusion is that homography in a text partly depends on the definition of part of speech, but that it derives from how the lexicon used is composed and the part of speech model then employed. Homograpy is probably also dependent on text genres. To the best of my knowledge, this field remains to be fully explored for Swedish.

9.2.1.2 Initial lexical material

The initial lexical materials may influence performance through its inventory of units and its information structure.

The units used in a lexicon describe what words are represented. A large lexicon therefore gives a better result than a small one. SMDB consists of approximately 1.2 million forms, thus providing extensive coverage. Since SMDB is used for annotation, an additional analysis of compounds is performed on unknown words, Berg et al. (2003). Remaining unknown words are subjected to morphological analysis in EPOS.

The information structure in the lexicon differs between heterographs and homographs. The frequency of homographs in a lexicon is also of great importance. This, in turn, has a great effect on performance, since the more tags a homograph might use, the more difficult the task of randomly choosing the correct tag will be. The figures compared in chapter 6 are 1.07 tags per word in a lexicon generated from SUC by Megyesi, versus 1.39 tags per word in SMDB. The logical conclusion to be drawn from these figures is that it is more difficult to tag a text correctly with SMDB than with the SUC-generated lexicon.

In chapter 5 a new way is proposed to compare regular performance in relation to how it is affected by homography in a lexicon. The result of using this new way of evaluation, $H_{measure}$, will be that performance figures will be upgraded according to increasing homography in a lexicon or a text.

9.2.1.3 Method

The method used in this work for the construction of EPOS is a relatively simple statistical method based on trigram. In certain especially problematic areas, some specific measures had to be taken, apart from the statistical analysis. A morphological analysis was also used for analysing unknown words. The statistical method has, in the two versions of EPOS, not proved crucial to performance, but the possibility cannot be ruled out that another statistical method would improve performance.

The problems in connection with disambiguation were mainly related to incorrectly tagged words in the textual material. Errors and problems in the tagging have been discussed in chapter 4. Errors will also be discussed in the conclusions about evaluation method in this summary. The errors previously discussed were, e.g., how spoken words should be handled.

SMDB is a lexicon created for annotating written text only. Therefore, a separate check of words which might be spoken is needed, or these words may be incorrectly tagged. Another problem was the unknown words, which might affect performance noticeably. Improvement of the module that handles unknown words is therefore needed. It should not only draw conclusions on the basis of morphological endings, but other contextual factors should also be taken into account. The most common unknown words are proper nouns, foreign words and nouns.

Improvement of performance with the method used is investigated in chapter 7. Common errors, along with possible solutions, are presented. Special measures have been taken to correct common errors, such as agreement analysis, analysis of the infinitive marker *att* ('to'), analysis of the word *som* ('which'), and complementary statistical analysis in the form of Bigram. Of these special measures the approach using the rule type Bigram proved the most successful. It generated only 3 per cent errors, to be compared with some general contextual analyses performed in EPOS, where little or no information was available when erroneously generated solutions could be as high as 39 per cent. These approaches are described in chapter 5.

Research within the area of POS tagging has proved that substantial improvement can only be achieved up to a certain point. After that the effect gets much smaller. The conclusions about different approaches to improvement of performance are that special solutions focused on problem areas are very effective and, further, that an extended version of agreement control will, at this stage, yield the best result.

9.2.2 Indirect factors

Tagset and choice of evaluation method are some of the factors which have an indirect effect on the measured performance.

9.2.2.1 Tagset composition

A tagset can affect a POS tagger's performance indirectly by its size and its contents.

In this thesis work, no investigations were carried out on the relation between tagset size and performance. On the other hand, the importance of how graphic words are represented is discussed. When comparing size and content in SUC's tagset and SMDB's tagset, there is an obvious difference in size, 153 and 133 tags, respectively. Further investigation showed that,

in spite of this, SMDB in many cases represents a more specific tagging particularly for adjectives. Therefore, there may be more tags in SMDB representing adjectives than there are in SUC's tagset. This gave rise to unexpected homography in the mapping process between the tagsets, since one tag in SUC's tagset would correspond to more than one tag in SMDB's tagset. Since SMDB's tagging then becomes more detailed, this proves that size does not determine how good or how detailed a tagging may be.

Other studies by, e.g. Megyesi, Tufis, Brants and Elworthy, which all aim at showing how tagset size affects performance, are accounted for in chapter 6. The obvious, unanimous conclusion of these studies is that a smaller tagset implies better performance.

When it is of minor importance to detect as much morphological information as possible in a text, the method of decreasing the tagset size when evaluating may be successful. But since the goal of tagging with EPOS is to detect as much syntactic and morphological information as possible, to promote further analysis, this is not a relevant option.

9.2.2.2 Evaluation method

The purpose of evaluation is to describe the performance of a process, in this case part of speech tagging.

The choice of evaluation method is important in order to account for the performance correctly and fairly. It is common to use superficial methods of evaluation, such as correctness, recall and precision. These measures do not give a detailed idea of the performance, but rather an overview of it, mostly with only one measure for all parts of speech.

In this study, a more detailed account of performance has been made concerning known and unknown words, but also with regard to each part of speech. Among others, Elworthy has carried out similar studies on known and unknown words.

Any evaluation must be performed and viewed in relation to its purpose. If the purpose is to lemmatize, identify new words or grammatically parse the tagged text, then the evaluation must reflect what is relevant to this purpose. The evaluation should not cover more than is needed for the purpose, since performance often suffers as increasingly specific information is being annotated in a text.

A survey of errors detected in tagging has been performed, where errors

and error types have been demonstrated in order to increase the possibilities of improving the performance of the tagger EPOS. Other error categories are also possible, e.g. by distinguishing between morphological errors and category errors; cf. Källgren (1992). This is discussed in chapter 2.

In chapter 2 and chapter 5 an overview is provided on commonly used measures for evaluating the performance of a tagger. Since homography has proved to be of great importance for performance, a proposal for a new measure has been given, $H_{measure}$. The purpose of this measure is to give a higher performance the more homography is present in a lexicon or a text and, conversely, a lower value when there is less homography present. Another measure is also mentioned which can be used to describe the difficulty of randomly disambiguating homographs with a varying number of tags.

My conclusion concerning evaluation is that in order to draw any firm conclusions about erroneous tagging and how a tagger's performance could be improved for some given conditions, a thorough survey of errors must be carried out. To find better solutions, it is also important to examine what types of disambiguation methods do not function well and to find out in what situations this is the case. In view of varying circumstances, it is important to evaluate as many relevant factors as possible. Therefore, evaluation measures such as the one proposed in this work should also be used when considering the size of a training corpus, the size and content of a tagset as well as other relevant factors making disambiguation more difficult.

9.3 Final conclusions

This study has shown that there are several factors which, directly or indirectly, affect the performance of a tagger. The computational method used in disambiguation did not seem to have more effect on performance than other factors. It appears that focusing on certain problem areas may help to improve performance the most at the present stage of development of EPOS. The method of evaluation is extremely important for how the performance of a tagger is viewed, but other measures designed to take relevant factors into consideration are also needed.

There is, so far, no POS tagger which tags texts 100 per cent correctly. As shown in this investigation, there are several areas in need of improvement. However, work on improvements is a time-consuming task; therefore, the work started in this study will be continued.

Litteraturreferenser

Aijmer K. and Altenberg B. (eds.) (1991), *English Corpus Linguistics*: Studies in Honour of Jan Svartvik, London, Longman.

Allén S. (1970, 1971), *Nusvensk frekvensordbok (NFO)*, Del 1-3, Data linguistica, Almqvist & Wiksell, Stockholm.

Andersson E. 1996, *Grammatik från grunden*, Hallgren och Fallgren studieförlag, Stockholm.

Andersson, Cooper, Hashemi, *A Finite-state Word Analyser*, Göteborg University, 1998.

Berg S., Samuelson K. (1988), *SAOL as a Spelling-Checker Dictionary*, i Studies in Computer-Aided Lexicology, Department of Computational Linguistics, Göteborg University.

Berg S., Cederholm Y. (2001), *Att hålla på formerna. Om framväxten av Svensk morfologisk databas*, I festskrift: Gäller stam, suffix och ord, Meijerbergs institut, Göteborg.

Berg S., Cederholm Y., Gellerstam M. (2003), *Svensk morfologisk databas baserad på tolfte upplagan av Svenska Akademiens ordlista*, under bearbetning.

Borin L. (2000), *Something Borrowed, Something Blue: Rule-Based Combination of POS Taggers*, i Second International Conference on Language Resources and Evaluation. Proceedings, Volume I, Athens, Greece.

Borin L. (2002), *Enhancing tagging performance by combining knowledge sources*, i Korpusar i forskning och undervisning. Rapport från ASLA:s höstsymposium, Växjö.

Brants, Thorsten (1998), *TnT - A Statistical Part-of-Speech Tagger*, Installation and User Guide. Saarland University, Computational Linguistics, (http://www.coli.uni-sb.de/~thorsten/tnt)

Brill E. (1993), A *Corpus-Based Approach to Language Learning*, A Dissertation in Department of Computer and Information Science, University of Pennsylvania.

Brill E. and Marcus M. (1992), *Tagging an unfamiliar text with minimal human supervision*, i Proceedings of the Fall Symposium on Probabilistic Approaches to Natural Language, American Association for Artificial Intelligence.

Brill E., (1994), *Some advances in transformation-based part-of-speech tagging*, i proceedings från AAAI'94.

Brill E., (1999), *Tagging Unknown Words*, in Syntactic wordclass tagging, ed. van Halteren H., Kluwer Academic Publishers, The Netherlands.

Brodda B., *Probems with Tagging - and a Solution.* Nordic Journal of Linguistics 5, s. 93-116, Scandinavian University Press.

Carlberger J., Kann V. (1999), *Implementing an efficient part-of-speech tagger*, Nada, Numerical Analysis and Computing Science, KTH, Stockholm.

Charniak E., *Statistical Language Learning*, The MIT Press, Cambridge, Massachusetts, London, England, 1993.

Chen, Bangalore, Vijay-Shanker, *New Models for Improving Supertag Disambiguation*, (Trigrams) I Proceedings of the 9th EACL, Bergen, Norway, s. 188

Church Kenneth W. 1988, *A stochastic parts program and noun phrase parser for unrestricted text.* I ANLP 2, s. 136-143.

Cutting D., Dupiec J., Pedersen J. and Sibun P. (1992), *A practical part-of-speech tagger*, i Proceedings of the Third Conference on Applied Natural Language Processing (ANLP-92), Trento, Italy, s.133-40.

Cutting D., 1993, *Porting a stochastic part-of-speech tagger to Swedish*, proceedings från Nodalida.

Daelemans W., 1999, Machine Learning Approaches, i van Halteren (ed.) Syntactic Wordclass Tagging, Kluwer Academic Publishers.

De Rose S. (1991:9), *An analysis of probabilistic grammatical tagging methods*, i Johansson and Stenström 1991, s. 9-13.

De Rose S. (1988, pp. 31-39), *Grammatical Category Disambiguation by Statistical Optimization.* Computational Linguistics, Vol. 14, Nr. 1.

EAGLES, Expert Advisory Groups on Language Engineering Standards, (1999), http://www.ilc.pi.cnr.it/EAGLES/home.html.

Ejerhed et al., *The Linguistic Annotation of the Stockholm-Umeå Corpus project, Technical Report No. 33*, Univ of Umeå, 1992.

El-Beze M., Merialdo B., (1999), *Hidden Markov Models*, i Syntactic Wordclass Tagging, s. 263-284, van Halteren H. (ed.), Kluwer Academic Publishers.

Elworthy D., (1995), *Tagset Design and Inflected Languages*, Sharp Laboratories of Europe Ltd, Oxford.

Fraces W.B, Baeza-Yates R. (1992), *Information Retrieval.* Data Structures & Algorithms. Prentice-Hall, New Jersey.

Francis W. N. and Kucera H. (1982), *Frequency Analysis of English* Usage: Lexicon and Grammar Boston: Houghton Mifflin.

Gambäck B., *Processing Swedish Sentences: a Unification-based Grammar*, Dissertation KTH, Stockholm, 1997.

Garside R. (1987:1), *The CLAWS word-tagging system*, i Garside et al. (1987:30-41).

Garside R., Leech G. and Sampson G. (eds.) (1987), *The Computational Analysis of English, A Corpus-Based Approach*, London, Longman.

Garside R., Leech G. and McEnery A. (eds.) (1997), *Corpus Annotation, Linguistic Information from Computer Text Corpora*, Addison Wesley Longman Inc., New York.

Gilbert, R. & Armstrong S., (1995), *Tagging tool. MULTEXT* Deliverable 2.4.1.

Greene B. B. and Rubin G. M. (1971), *Automated Grammatical Tagging of English Providence*, RI: Department of Computer Science, Brown University.

Grossman D. A., Frieder O. (1998), *Information Retrieval.* Algorithms and heuristics. Kluwer Academic Publishers, The Netherlands.

Halliday M. A. K. (1991), *Corpus studies and probabilistic grammar*, i Aijmer and Altenberg (1991: 30-40).

Hassler-Göransson C. (1966), *Ordfrekvenser i nusvenskt skriftspråk*, Skriptor, Stockholm.

Johansson S., Stenström A.-B. (eds), *English Computer Corpora*: Selected Papers and Research Guide, Berlin: Mouton de Gruyter, 1991.

Johansson Kokkinakis S., Kokkinakis D., *Några problem vid automatisk analys av text*, Festskrift, Inst. f. svenska språket, Göteborgs universitet, 1999.

Jurafsky D., Martin J.H. (2000), *Speech and Language Processing*, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall, New Jersey.

Karlsson F. (1992), SWETWOL: *A Comprehensive Morphological Analyser for Swedish*. Nordic Journal of Linguistics, 15, s. 1-45.

Kokkinakis D, Johansson Kokkinakis S. (1997), *A Robust and Modularized Lemmatizer/Tagger for Swedish Based on Large Lexical Resources*, Inst. f. svenska språket, Göteborgs universitet.

Kokkinakis D., Johansson Kokkinakis S. (1999), *A Cascaded Finite-State Parser for Syntactic Analysis of Swedish*, In Proceedings of the 9th EACL, s. 245-248, Bergen, Norway.

Källgren G. (1992), *Making maximal use of surface criteria in large-scale parsing: the MorP parser*, Papers from the Institute of Linguistics, University of Stockholm, Publication 60, Institute of Linguistics, Stockholms universitet.

Källgren G. (1998), *Documentation of the Stockholm-Umeå Corpus (DRAFT)*, Stockholms universitet.

Lager T. (1999), *The μ-TBL System: Logic Programming Tools for Transformation-Based Learning*, i Proceedings of Third International Workshop on Computational Natural Language Learning (CoNLL-99), Bergen.

Ljung M., Ohlander S. (1987), *Allmän grammatik*, Liber, Stockholm.

Leech G., Smith N., (1999), *The Use of Tagging*, i van Halteren (ed.), *Syntactic Wordclass Tagging*, s. 23-36, Kluwer Academic Publishers.

Leech G., Wilson A., (1999), *Standards for Tagsets*, i van Halteren (ed.), *Syntactic Wordclass Tagging*, s. 55-80, Kluwer Academic Publishers.

Malmberg B., (1962), *Nya vägar inom språkforskningen: en orientering i modern lingvistik*, Scandinavian university books. Stockholm.

Manning C. D., Schütze H., (1999), *Foundations of statistical natural language processing*. The MIT Press, Cambridge, Massachusetts, London, England.

Marcus M., Santorini B., Marcinkiewicz M.A. (1993). *Building a large annotated corpus of English: the Penn Treebank*. Computational Linguistics, 19:2. s. 313-330.

Marshal I., (1987). *Tag selection using probabilistic methods*. I Garside et al. (eds.), *The computational analysis of English: a corpus-based approach*, s. 42-56, London, Longman.

McEnery T. and Wilson A. (1997), *Corpus Linguistics*, Edingburgh University Press, Edingburgh.

Megyesi, B. (2001), *Comparing Data-Driven Learning Algorithms for PoS Tagging of Swedish*, Department of Speech, Music and Hearing, KTH, Stockholm.

Megyesi, B. (2002), *Data-Driven Syntactic Analysis. Methods and Applications for Swedish*, Doctoral Dissertation, Department of Speech, Music and Hearing, KTH, Stockholm.

Mitchell, T. (1997), *Machine Learning*, International Editions, Singapore, 1997.

Nagata, M. (1994), *A stochastic Japanese morphological analyser using a Forward-DP Backward-A*N-Best search algorithm*. In Proceedings of COLING-94, Kyoto. s 201-207.

Orphanos et. al (1999), *POS Disambiguation and Unknown Word Guessing with Decision Trees*, In Proceedings of the 9th EACL, s. 134-141, Bergen, Norway.

Prütz K. (1999), *Sammanställning av en träningskorpus på svenska för träning av ett automatiskt ordklasstaggningssystem*. Reports from the ETAP project. ISSN 1401-923X Ed. Sågvall Hein A., Uppsala universitet.

Prütz K. (2002), *Part-of-speech tagging for Swedish, in Parallel corpora, parallel worlds*, Language and Computers Vol. 43, ed. Borin L., Rodopi.

Ratnaparkhi A. (1996), *A Maximum Entropy Model for Part-Of-Speech Tagging*, I proceedings från ACL-SIGDAT Conference on Empirical Methods in Natural Language Processing, Philadelphia, Pennsylvania.

Ratnaparkhi A. (1999), *Natural Language Learning with the Maximum Entropy Framework*, EACL'99, Bergen, Norway.

Schmid H., (1994), *Part-of-speech tagging with neural networks*. I Proceedings of COLING-94, Kyoto. s. 172-176.

Shannon C. and Weaver W. (1949), *The Mathematical Theory of Communication*. Urbana, IL, University of Illinois Press.

Skut W., Krenn B., Brants T., Uszkoreit H. (1997), *An annotation scheme for free word order languages*. I Proceedings of the fifth conference on applied natural language processing ANLP-97, Washington, DC.

Soderland S., Fisher D., Lehnert W., (1997), *Automatically Learned vs. Hand-crafted Text Analysis Rules*, CIIR Technical Report (TC-44). Natural Language Processing Laboratory, University of Massachusetts.

Svenska Akademien (1999), *Svenska akademiens grammatik (SAG)*, MediaPrint, Uddevalla 1999.

Tufis D., (2000), *Principled Hidden Tagset Design for Tiered Tagging of Hungarian*, i proceedings LREC 2000, Athens, s.1421-1426.

van Guilder L. (1995), *Automated Part of Speech Tagging: A Brief Overview*, Georgetown University.

van Halteren H. (ed.) (1999), *Syntactic Wordclass Tagging*, Kluwer Academic Publishers. Printed in the Netherlands.

van Rijsbergen C.J. (1979), *Information Retrieval*, Butterworths, London. (http://www.dcs.gla.ac.uk/Keith/Preface.html).

Voutilainien A, *Does Tagging Help Parsing? A Case Study on Finite State Parsing*, In Proceedings FSMNLP '98, Ankara, Turkey, 1998.

Voutilainen A., 1999, *A Short History of Tagging*, in Syntactic Wordclass Tagging s. 9-21, H. van Halteren (ed.), Kluwer Academic Publishers. Printed in the Netherlands.

Weischedel, R., Meteer M., Schwarts R., Ramshaw L., Palmuzzi J. (1993), *Coping with ambiguity and unknown words through probabilistic models.* Computational Linguistics, 19:2.

Zavrel J., Daelemans W., (1999), *Recent Advances in Memory-Based Part-of-Speech Tagging*, Tilburg University.

Åström M., (1995), A probabilistic tagger for Swedish using the SUC tag-set. *In Proceedings of the Conference on Lexicon + Text.* Lexicographica - Series Maior. Tübingen: Niemeyer.

APPENDIX A. Taggуppsättning för EPOS 1.0

EPOS 1.0		
A0(CS)00GD	A0P00N0	AP00(SP)N(ID)
A0(CS)00ND	A0P0PN0	AP00000
A00(NU)(SP)0(ID)	A0P0SGD	AP000N0
A00(NU)(SP)G(ID)	A0P0SND	AP00PN0
A00(NU)(SP)GD	A0P0SNI	AP00SN0
A00(NU)(SP)N(ID)	A0PMSG(ID)	AP00SNI
A00(NU)(SP)ND	A0PMSGD	AP0MSGI
A00(NU)PG0	A0PMSGI	AP0MSNI
A00(NU)PN0	A0PMSN(ID)	AP0N(SP)G(ID)
A00(NU)SG0	A0PMSND	AP0N(SP)N(ID)
A00(NU)SN0	A0PMSNI	AP0NSGI
A000000	A0PN000	AP0NSNI
A0C(NU)(SP)0(ID)	A0PNS0I	AP0USGI
A0C(NU)(SP)00	A0PNSGD	AP0USNI
A0C(NU)(SP)G(ID)	A0PNSGI	APP0000
A0C(NU)(SP)G0	A0PNSNI	C
A0C(NU)(SP)GD	A0PU000	CA
A0C(NU)(SP)N(ID)	A0PUS0I	CC
A0C(NU)(SP)N0	A0PUSGI	CI
A0C(NU)(SP)ND	A0PUSNI	CO
A0C(NU)PGI	A0S(NU)(SP)00	CS
A0C(NU)PNI	A0S(NU)(SP)0D	CT
A0C00GD	A0S(NU)(SP)GD	CU
A0C00GI	A0S(NU)(SP)ND	DF0P0
A0C00N0	A0S(NU)0G0	DF0S0
A0C00ND	A0S(NU)0GD	DFNS0
A0C00NI	A0S(NU)0N0	DFUS0
A0P(NU)(SP)0(ID)	A0S(NU)0ND	DINS0
A0P(NU)(SP)G(ID)	A0S(NU)PGD	DIUSN
A0P(NU)(SP)GD	A0S(NU)PND	DIUS0
A0P(NU)P0D	A0S(NU)SGI	F
A0P(NU)PG(ID)	A0S(NU)SNI	I
A0P(NU)PGD	A0S0(SP)GD	MC
A0P(NU)PGI	A0S0(SP)ND	MO
A0P(NU)PND	A0S000D	NC(NU)(SP)GI
A0P(NU)PNI	A0S00N0	NC(NU)(SP)NI
A0P(NU)S0I	A0S0SN0	NC(NU)000
A0P(NU)SGI	AP0(NU)(SP)(NG)(ID)	NC(NU)PGD
A0P(NU)SNI	AP0(NU)(SP)0(ID)	NC(NU)PGI

Tabell 71. Taggуppsättning för EPOS 1.0.

NC0(SP)(NG)I	NCNPGD	P(PR)0SN00
NC0(SP)GD	NCNPGI	P(PR)NP(NU)0O
NC0(SP)GI	NCNPN0	P(PS)0PN0O
NC0(SP)ND	NCNPND	P(PS)NPNDO
NC0(SP)NI	NCNPNI	P(PS)NPNIO
NC0000	NCNS(NG)I	P(PS)UPNDO
NC000I	NCNS00	P(SX)0PN0O
NC00G0	NCNSG(ID)	P(SX)UPNIO
NC00GD	NCNSGD	P(TR)0
NC00GI	NCNSGI	P(TR)0(SP)GD0
NC00N0	NCNSND	P(TR)0(SP)ND0
NC00ND	NCNSNI	P(TR)NSGI0
NC00NI	NCU(SP)(NG)I	P(TR)NSNI0
NC0PG(ID)	NCU(SP)GD	P(TR)USGI0
NC0PG0	NCU(SP)GI	P(TR)USNI0
NC0PGD	NCU0(NG)I	P0(NU)PN(ID)0
NC0PGI	NCU0GD	P0NSNI0
NC0PN(ID)	NCU0GI	P0USN
NC0PN0	NCU0ND	PC0PG00
NC0PND	NCU0NI	PC0PN00
NC0PNI	NCUPG(ID)	PD0
NC0S0(ID)	NCUPG0	PD00N
NC0SG0	NCUPGD	PD0PN
NC0SGD	NCUPGI	PD0SGI0
NC0SGI	NCUPN0	PD0SNI0
NC0SN0	NCUPND	PDMSGI0
NC0SND	NCUPNI	PDMSN
NC0SNI	NCUS(NG)I	PDNSGI0
NCMSGI	NCUS0I	PDNSN
NCMSNI	NCUSG(ID)	PDNSNI0
NCN(SG)GI	NCUSG0	PDUPGD0
NCN(SP)(NG)I	NCUSGD	PDUPND0
NCN(SP)0I	NCUSGI	PDUSGI0
NCN(SP)G(ID)	NCUSN0	PDUSN
NCN(SP)GD	NCUSND	PDUSNI0
NCN(SP)GI	NCUSNI	PI0PN
NCN0GD	NP	PI0SN
NCN0GI	NPG	PIMSN
NCN0ND	NTB	PP(NU)PGI0
NCN0NI	P(PR)(NU)PG00	PP(NU)PNI0
NCNP0D	P(PR)(NU)PN00	PP(NUM)SNI0
NCNPG(ID)	P(PR)(NU)PN0O	PP0
NCNPG0	P(PR)(NU)SG00	PP0PN

Tabell 71. Tagguppsättning för EPOS 1.0.

PP0PN00	RJ0	VMU0A
PP0SN00	RM0	VMU0P
PPFS00O	RN0	X
PPFSG00	RP0	Y
PPFSN	RT0	
PPFSNI0	S	
PPMS00O	TT	
PPMSG00	V(MA)ISA	
PPMSGD0	V(MA)N0A	
PPMSN	V(MA)U0A	
PPNPN	V000A	
PPNSGI0	V0IPA	
PPNSN	V0IPP	
PPNSNI0	V0ISA	
PPU(SP)GD0	V0ISP	
PPU(SP)ND0	V0N0A	
PPUPG00	V0U0A	
PPUPN00	V0U0P	
PPUSGI0	VAI0A	
PPUSN	VAIPA	
PPUSNI0	VAU0A	
PR0	VM(IN)00	
PRNSN	VM(NI)0P	
PRUSN	VM000	
PS(NU)PN00	VM00A	
PS00N	VM00P	
PSNSN00	VM0PP	
PSUSN	VM0SP	
PSUSN00	VM0UP	
PT0	VMI00	
PX(NU)(SP)00O	VMI0A	
PX00N	VMIPA	
PX0S00O	VMIPP	
PX0SN0O	VMIS0	
R	VMISA	
RA0	VMISP	
RD0	VMM0A	
RDB	VMN0A	
RDC	VMN0P	
RDS	VMS00	
RE0	VMSPA	
RF0	VMSSA	
RI0	VMU00	

Tabell 71. Tagguppsättning för EPOS 1.0.

APPENDIX B. Taggуppsättning för SMDB/EPOS 2.0

tagg	förklaring
AF0*PG*	particip (perfekt) utr/neu plu gen indef/def
AF0*PN*	particip (perfekt) utr/neu plu nom indef/def
AF0*SGD	particip (perfekt) utr/neu sing gen def
AF0*SND	particip (perfekt) utr/neu sing nom def
AF0MSGD	particip (perfekt) maskulinum singular genitiv bestämd
AF0MSND	particip (perfekt) maskulinum singular nominativ bestämd
AF0NSNI	particip (perfekt) neu sing nom indef
AF0USGI	particip (perfekt) utrum singular genitiv obestämd
AF0USNI	particip (perfekt) utrum singular nominativ obestämd
AP0**G*	particip (presens) utrum/neutrum singular/plural genitiv obestämd/bestämd
AP0**N*	particip (presens) utrum/neutrum singular/plural nominativ obestämd/bestämd
AQC****	
AQC**G*	adjektiv komparativ utrum/neutrum singular/plural genitiv obestämd/bestämd
AQC**N*	adjektiv komparativ utrum/neutrum singular/plural nominativ obestämd/bestämd
AQP****	
AQP**N*	adjektiv positiv utrum/neutrum singular/plural nominativ obestämd/bestämd
AQP**NI	adjektiv positiv utrum/neutrum singular/plural nominativ obestämd
AQP*PG*	adjektiv positiv utrum/neutrum plural genitiv obestämd/bestämd
AQP*PN*	adjektiv positiv utrum/neutrum plural nominativ obestämd/bestämd
AQP*PNI	adjektiv positiv utrum/neutrum plural nominativ obestämd
AQP*SGD	adjektiv positiv utrum/neutrum singular genitiv bestämd
AQP*SND	adjektiv positiv utrum/neutrum singular nominativ bestämd
AQPMSGD	adjektiv pos mask sing gen def
AQPMSND	adjektiv pos mask sing nom def
AQPNSGI	adjektiv pos neu sing gen indef
AQPNSN*	adjektiv pos neu sing nom indef/def
AQPNSNI	adjektiv pos neu sing nom indef
AQPU***	
AQPUSGI	adjektiv pos utr sing gen indef
AQPUSN*	adjektiv pos utr sing nom indef/def

Tabell 72. SMDB – taggуppsättning

tagg	förklaring
AQPUSNI	adjektiv pos utr sing nom indef
AQS**ND	adjektiv pos utr/neu sing/plu nom def
AQS**NI	adjektiv pos utr/neu sing/plu nom indef
AQS*PND	adjektiv pos utr/neu plu nom def
AQS*PNI	adjektiv pos utr/neu plu nom indef
AQSNSGD	adjektiv pos neu sing gen def
AQSNSND	adjektiv pos neu sing nom def
CC	konjunktion
CS	subjunktion
D0NS	
D0US	
DF*P	
DF*P	determinerare def utr/neu plur
DF*S	
DFMS	determinerare def mask sing
DFNS	determinerare def neu sing
DFUS	determinerare def utr sing
DH*P	pronomen (frågande)
DHNS	pronomen (frågande)
DHUS	pronomen (frågande)
DI**	determinerare indef utr/neu sing/plu
DI*P	determinerare indef utr/neu plu
DI*S	determinerare indef utr/neu sing
DIMS	determinerare indef mask sing
DINS	determinerare indef neu sing
DIUS	determinerare indef utr sing
FC	interpunktion (omgärdande)
FE	interpunktion (meningsavgränsande)
FI	interpunktion (inuti mening)
I	interjektion
MC*0N*	
MC*PG*	räkneord (grundtal) utrum/neutrum genitiv obestämd/bestämd
MC*PN*	räkneord (grundtal) utrum/neutrum genitiv obestämd/bestämd

Tabell 72. SMDB – taggуppsättning

tagg	förklaring
MC*PN*	räkneord (grundtal) utrum/neutrum nominativ obestämd/bestämd
MCNSNI	räkneord (grundtal) neutrum singular nominativ obestämd
MCUSNI	räkneord (grundtal) utrum singular nominativ obestämd
MO*SG*	räkneord (ordningstal) utrum/neutrum genitiv obestämd/bestämd
MO*SN*	räkneord (ordningstal) utrum/neutrum nominativ obestämd/bestämd
MOMSG*	räkneord (ordningstal) maskulinum singular genitiv obestämd/bestämd
MOMSN*	räkneord (ordningstal) maskulinum singular nominativ obestämd/bestämd
NC*000	
NCN***	
NCNPGD	substantiv neutrum plural genitiv bestämd
NCNPGI	substantiv neutrum plural genitiv obestämd
NCNPND	substantiv neutrum plural nominativ bestämd
NCNPNI	substantiv neutrum plural nominativ obestämd
NCNSGD	substantiv neutrum singular genitiv bestämd
NCNSGI	substantiv neutrum singular genitiv obestämd
NCNSND	substantiv neutrum singular nominativ bestämd
NCNSNI	substantiv neutrum singular nominativ obestämd
NCU***	
NCUPGD	substantiv utrum plural genitiv bestämd
NCUPGI	substantiv utrum plural genitiv obestämd
NCUPND	substantiv utrum plural nominativ bestämd
NCUPNI	substantiv utrum plural nominativ obestämd
NCUSGD	substantiv utrum singular genitiv bestämd
NCUSGI	substantiv utrum singular genitiv obestämd
NCUSND	substantiv utrum singular nominativ bestämd
NCUSNI	substantiv utrum singular nominativ obestämd
NP*SGD	proprium genitiv
NP*SND	proprium nominativ
PE000	pronomen (frågande)
PF**O	pronomen (personligt) def utr/neu sing/plu obj
PF*P*	pronomen (personligt) def utr/neu plu subj/obj
PF*PO	pronomen (personligt) def utr/neu plu obj
PF*PS	pronomen (personligt) def utr/neu plu subj

Tabell 72. SMDB – taggуppsättning

tagg	förklaring
PFMS*	pronomen (personligt) def mask sing subj/obj
PFNS*	pronomen (personligt) def neu sing subj/obj
PFUPO	pronomen (personligt) def utr plu obj
PFUPS	pronomen (personligt) def utr plu subj
PFUS*	pronomen (personligt) def utr sing subj/obj
PFUSO	pronomen (personligt) def utr sing obj
PFUSS	pronomen (personligt) def utr sing subj
PH**0	pronomen (frågande, relativt) utr/neu sing/plu
PH*P0	pronomen (frågande, relativt) utr/neu plu
PHNS0	
PHNS0	pronomen (frågande, relativt) neu sing
PHUS0	pronomen (frågande, relativt) utr sing
PI*P*	pronomen (personligt) indef utr/neu plu subj/obj
PINS*	pronomen (personligt) indef neu sing subj/obj
PIUS*	pronomen (personligt) indef utr sing subj/obj
PIUSS	pronomen (personligt) indef utr sing subj
PS***	pronomen (possesivt) utr/neu sing/plu subj/obj
PS*P*	pronomen (possesivt) utr/neu plu subj/obj
PSNS*	pronomen (possesivt) neu sing subj/obj
PSUS*	pronomen (possesivt) utr sing subj/obj
R0	
R0	adverb (oböjligt)
RC	adverb komparativ
RP	adverb positiv
RS	adverb superlativ
S	preposition
V****	
V0IPA	verb indikativ presens aktiv (V2)
V0IPP	verb indikativ presens passiv (V9)
V0ISA	verb indikativ preteritum aktiv (V3)
V0ISP	verb indikativ preteritum passiv (V10)
V0M0A	verb imperativ aktiv (V17)
V0M0P	verb imperativ passiv

Tabell 72. SMDB – tagguppsättning

tagg	**förklaring**
V0N0A	verb infinitiv aktiv (V1)
V0N0P	verb infinitiv passiv (V8)
V0SPA	verb konjunktiv presens aktiv
V0SPP	verb konjunktiv preteritum passiv
V0SSA	verb konjunktiv preteritum aktiv (V99)
V0U0A	verb supinum aktiv (V4)
V0U0P	verb supinum passiv (V11)
XF	annat språk
Y	förkortning
Z	infinitivmärke

Tabell 72. SMDB – taggуppsättning