



UPPSALA
UNIVERSITET

Institutionen för lingvistik och filologi
Språkteknologiprogrammet
Examensarbete i datorlingvistik
26 augusti 2005

Automatisk morfologisk analys av ungerska substantiv med PC-KIMMO

Jennie Gadeborg

Handledare:
Beáta Megyesi, Uppsala universitet

Sammandrag

Denna uppsats beskriver automatisk morfologisk segmentering och analys av ungerska substantiv. Ett program som gör automatisk morfologisk analys på ungerska substantiv har skapats. Programmet analyserar automatiskt ungerska substantiv utifrån deras morfologiska beståndsdelar. Skalprogrammet PC-KIMMO har använts, där har ordstammar, suffix och en uppsättning regler som täcker fonologiska oregelbundenheter i språket lagts till. Programmet har utvärderats på en femhundra ord stor testsvit, plockad ur "Hungarian National Corpus". Resultatet visar att 73,6% av orden fått korrekt analys, varav 11% fått mer än en analys och den korrekta analysen finns med som ett av alternativen.

Abstract

This paper describes automatic morphological segmentation and analysis of Hungarian nouns. A program for automatic morphological analysis of Hungarian nouns was created. PC-KIMMO which works as a shell was used. Stems, suffixes and rules were added to PC-KIMMO. The program has been evaluated with 500 random nouns selected from the "Hungarian National Corpus". 73,6% of the words were analyzed correctly. 11% of the correctly analyzed words got two or three analyses and amongst them were the correct analysis.

Innehåll

Sammandrag	ii
Abstract	iii
Tack	v
1 Inledning	1
1.1 Syfte	1
1.2 Uppsatsens upplägg	2
2 Ungerska substantivs fonologi och morfologi	3
2.1 Fonologi	3
2.2 Ungerska substantivs morfologi och morfotax	5
3 Automatisk morfologisk segmentering och analys	10
3.1 PC-KIMMO	12
3.2 Humor	13
3.3 Xerox morfologisk analys, tokenisering och disambiguering	13
4 Automatisk morfologisk segmentering och analys av ungerska substantiv	15
4.1 Data	15
4.2 Programmets utformning	15
5 Utvärdering	19
5.1 Utvärderingsmetod	19
5.2 Resultat av utvärderingen	19
5.2.1 Felanalyserade ord	23
5.2.2 Ej analyserade ord	23
6 Diskussion	24
7 Sammanfattning	26
Litteraturförteckning	27
A Kod	29
B Tagguppsättning	38

Tack

Jag vill först och främst tacka min handledare Beáta Megyesi vid Institutionen för lingvistik och filologi för all hjälp under uppsatsskrivandet. Jag vill också tacka Mats Dahllöf vid Institutionen för lingvistik och filologi för givande seminarier och Per Starbäck vid Institutionen för lingvistik och filologi för teknisk support och latex-mallen.

Nagy köszönet Pajzs Júliának, Oravecz Csabának, Sass Bálintnak (Magyar Tudományos Akadémia) és Alexin Zoltánnak (Szegedi Tudományegyetem) hogy segítettek nekem magyar szövegtárákat találni.

Tack till Marie Fryer Hydfors och Lina Andersson, mina personliga svenskfröknar, samt till Jesper Lundgren som underlättade mitt liv med det lilla skriptet. Tack till Mária Dugántsy och Vanda Czifra som lärt mig allt jag vet om ungerska. Jag vill också tacka Lars Borin vid Institutionen för svenska språket, Göteborgs universitet, för hjälpen jag fått när PC-KIMMO inte fungerat som jag önskat.

Slutligen vill jag också tacka Malin Gadeborg och Inger Gadeborg för det stöd de gett mig under arbetet med denna uppsats.

1 Inledning

Den del inom grammatiken som handlar om hur det ser ut inom ett ord kallas morfologi. Inom morfologi tittar man på hur ord byggs upp, det vill säga vilka delar ett ord får bestå av. Morfologisk segmentering innebär att man delar upp ett ord i stam och suffix (ändelser). Vid morfologisk analys får de olika delarna etiketter, man sätter olika namn på suffixen beroende vilka olika kategorier de tillhör, kategorierna kan vara plural, dåtid mm. Morfologisk segmentering och analys av ord är en viktig del inom många datorlingvistiska tillämpningar. En datorlingvistisk tillämpning innebär att någon form av språkbehandling görs med hjälp av ett program. Nedan följer en uppräknig på olika slags datorlingvistiska tillämpningar där morfologisk segmentering och analys används. Inom maskinöversättning kan segmentering användas för att skapa elektroniska ordlistor. Morfologisk analys av ord kan användas i en ordklassstaggare, där ett ord tilldelas en viss ordklass beroende på vilka suffix ordet har. Segmentering används inom ordbehandling för att skapa effektiva rättstavnings- och avstavningsprogram. I text-till-tal-system används morfologisk analys för att ge bättre uttal.

I ungerskan finns väldigt många böjningsformer för varje ord. Sådant som man exempelvis i många andra språk uttrycker med prepositioner uttrycks i ungerskan med kasussuffix. Eftersom ungerskan är så formrik blir morfologisk segmentering och analys extra viktigt. För att skapa till exempel en elektronisk ordlista för ungerska ord krävs någon form av morfologisk analysator för att inte ordlistan ska bli enorm. Ungerska ord kan innehålla många suffix och att lägga till alla varianter av ett ord är både tids- och utrymmeskrävande. Man skapar därför en lista som innehåller ordstammarna och en som innehåller de enskilda suffixen. Vidare behövs också ett program som gör den morfologiska analysen. Denna uppsats beskriver ett sådant program som automatiskt analyserar ungerska substantiv. Det program som skapats är skrivet i PC-KIMMO, ett skalprogram där lexikon och regler för ungerska substantiv har lagts till.

1.1 Syfte

Syftet med arbetet är att skapa en analysator som automatiskt analyserar ungerska substantiv och delar upp dem i deras morfologiska beståndsdelar samt markerar dessa på ett lättläst sätt. Programmet analyserar samtliga böjningssuffix som ungerska substantiv kan ha. Däremot lämnas avledningssuffixen oanalyzerade. De suffix som markeras är possessivsuffix, numerus och kasussuffixen: ackusativ, inessiv, elativ, illativ, superessiv, delativ, sublativ, adessiv, ablativ, allativ, terminativ, dativ,

essiv-modal, essiv-formalis, translativ-faktiv, instrumental-komitativ, instrumental-sociativ, kausal-final, distributiv, temporal, distributiv-temporalis och multiplikativ. Programmets korrekthet har utvärderats.

1.2 Uppsatsens upplägg

Uppsatsen består av fyra huvuddelar. I den första delen presenteras ungerska substantivs grammatik och automatisk morfologisk segmentering och analys. I den andra delen finns kapitel 4, Automatisk morfologisk segmentering och analys av ungerska substantiv, där beskrivs programmet som skapats. I den tredje delen finns utvärderingskapitlet, kapitel 5, som presenterar hur utvärderingen av programmet gjorts samt resultatet av denna. Den sista delen består av diskussionen med några ord om framtidsutsikter i kapitel 6 och en sammanfattning i kapitel 7.

2 Ungerska substantivs fonologi och morfologi

Ungerskan tillhör den finsk-ugriska språkfamiljen. Ungerska talas av ca. 15 miljoner människor. Ordförrådet består av ord med finsk-ugriskt ursprung men också låneord från många språk bland annat tyska och turkiska. Karaktäristiska drag hos ungerskan är att språket är mycket formrikt, suffix läggs efter varandra för att bilda avledningar och nya böjningsformer, se exempel 5. Ungerskan har vokalharmoni vilket innebär att ord innehåller antingen endast främre eller endast bakre vokaler. En stam med främre vokaler kan också bara få suffix med främre vokaler, se exempel 2. Gränserna mellan ordklasserna är inte fasta. Samma ordstam kan användas för att bilda både substantiv, adjektiv, adverb och verb. Ett exempel är *meleg* som betyder både 'värme' och 'varm'. Ungerskan har fri ordning för orden i satsen. Ordföljden baseras på informationsstrukturen. Det vill säga, vilket ord som kommer först i satsen beror på vad som är viktigast. Ny information hamnar först i satsen. Meningen 'Jag har en röd boll' kan översättas med *Van piros labdám* 'finns röd boll min' om det som är viktigt är att jag har en boll. Om jag däremot vill påpeka att bollen är röd översätts meningen med *Piros labdám van* 'röd boll min finns'.

Följande avsnitt tar upp företeelser inom ungersk grammatik som är aktuella för mitt program. Avsnittet är alltså inte tänkt som en uttömmande deskriptiv grammatik. I denna inledande del har jag utgått från Dugántsy (1997) och Megyesi (1998).

2.1 Fonologi

Det ungerska alfabetet består av 44 bokstäver, se exempel 1, och kallas det fullständiga ungerska alfabetet (Storlind, 2002). Vokalerna är indelade i par med en lång och en kort variant. De långa varianterna av varje vokal markeras med akut eller dubbelakut accent till exempel "é, é" och "ö, ő".

Exempel 1 Det ungerska alfabetet

a á b c cs d dz e é f g gy h i í j k l ly m n ny o ó ö ő p q r s sz t ty u ú ü ű v w x
y z zs

Ungerskan har, som tidigare nämnts, vokalharmoni vilket, enligt traditionell ungersk grammatik, innebär att alla ingående vokaler i ett ord antingen är främre eller bakre.

Till de främre vokalerna räknas "i, í, ü, ű, e, é, ö, ő". Till de bakre vokalerna räknas "u, ú, o, ó, a, á". Vokalerna "i, e" har en specialställning och kan förekomma tillsammans med de bakre vokalerna, se exempel 2, se appendix B för förklaringar till de morfologiska kategorierna. Vokalharmonin är viktig för suffixen. Det finns oftast minst två varianter av varje suffix, en variant med främre vokaler och en med bakre, eftersom vokalharmonin omfattar suffixsystemet. När ord slutar på en konsonant läggs en bindevokal mellan ordet och suffixet om suffixet börjar på en konsonant. Bindevokalerna omfattas också av vokalharmonin. Till ett ord med främre vokaler väljs en främre bindevokal. I sammansatta ord kan främre och bakre vokaler förekomma tillsammans, i sådana ord rättar sig suffixen efter den senare delen av det sammansatta ordet (Lavotha och Lavotha, 1973).

Exempel 2 Vokalharmonin

esküvő-k-ön	esküvő-PL-SUP	'på bröllop+PL'	endast främre vokaler
karácsony-kor	karácsony-TEM	'vid jul'	endast bakre vokaler
diák-ok-kal	diák-PL-INS	'med studenter'	"i" med bakre vokaler
eladó-hoz	eladó-ALL	'till försäljaren'	"e" med bakre vokaler

Det finns flera typer av assimilation i ungerskan. De markeras vanligtvis inte i skrift. Det finns dock undantag. Ett exempel är total assimilation med morfologisk bakgrund. Den påverkar substantiv och markeras i skrift. Den är progressiv och förekommer vid "v". "v" i början av suffix assimileras till ordstammens slutkonsonant, se exempel 3.

Exempel 3 Assimilation

állat	val =>	állat-tal
djur:NOM	INS	djur-INS
'djur'	'med'	'med djur'

Mellan stam och suffix eller mellan olika suffix används ibland bindevokaler. Vilka bindevokaler som är aktuella för vilket suffix anges i den detaljerade listan av suffix, senare i detta kapitel. Alla stammar som slutar på "a" eller "e" får en förlängd sista vokal när suffix fogas till ordet, se exempel 4.

Exempel 4 Vokalförlängning

szoba	ban =>	egy szobá-ban
rum:NOM	INE	ett rum-INE
'rum'	'i'	'i ett rum'

2.2 Ungerska substantivs morfologi och morfotax

Ungerskan är ett agglutinerande språk med flekterande drag. Det innebär att de flesta grammatiska relationerna, men inte alla, uttrycks med hjälp av olika affix. I agglutinerande språk ordnas de bundna formerna som ”pärlor på en tråd” (Sproat, 1992), det vill säga på en rad efter varandra. Stammen och suffixen är alltså tydligt skilda åt. Varje affix har sin bestämda plats och morfemen konkateneras efter varandra i tur och ordning, se exempel 5. Varje affix är ett unikt morfem som representerar en egen-skap. De flekterande dragen visar sig inte bland substantiven och förklaras därför inte närmre här.

Exempel 5 Ordbildning

tüdő-gyulladás-a-tól lung-inflammation-PSe3-ABL 'av hans/hennes lunginflammation'

I bland annat svenskan skiljer man på avledningssuffix och böjningssuffix. Avledningssuffixen förändrar ordens betydelse. Böjningssuffixen har grammatiska funktioner. Enligt den traditionella ungerska grammatiken delar man in ungerskans suffix i tre delar. Det finns avledningssuffix som kallas ”képző” och ändrar ordens betydelse. Böjningssuffixen delas in i två delar ”jel” och ”rag”. Både ”jel” och ”rag” uttrycker grammatiska relationer, se exempel 6. Bland de suffix som kategoriseras under benämningen ”rag” återfinns kasussuffixen och verbens personsuffix. Övriga böjningssuffix är av typen ”jel”. Denna indelning i tre delar görs inte på grund av vad suffixen uttrycker utan beroende på i vilken ordning de olika typerna av suffix får förekomma i ordet. De olika suffixsorterna har bestämd ordning i ord. Direkt efter rotmorfemet kommer avledningssuffixen. De kan vara flera till antalet. De följs av en ”jel” som i sin tur följs av en ”rag”, se exempel 7. Morfemordningen är alltså stam samt ett eller flera avledningssuffix som följs av:

- numerus + possessivmärke + kasus (se exempel 8)
- numerus + possessivsuffix + kasus (se exempel 9)

Exempel 6 Substantivens suffixtyper

képző	jel	rag
asztal-os	asztal-ok	az asztal-on
bord-KÉPZŐ	bord-PL	bordet-SUP
'snickare'	'bord+PL'	'på bordet'

Exempel 7 Suffixens ordning

az asztal-os-ok-on bord-KÉPZŐ-JEL-RAG 'på snickare+PL'
--

Exempel 8 numerus+possessivmärke+kasus

asztal-ok-é-t asztal-PL-POS-ACC 'bordens-ACC'

Exempel 9 numerus+possessivsuffix+kasus

fá-i-m-on träd-PL-PSeIi-SUP 'på mina träd'
--

Nedan följer en mer detaljerad genomgång av substantivsuffixen. Först presenteras avledningssuffixen och sedan följer de suffix som tillhör kategorin jel:

- Avledningssuffix

I ungerskan finns en rad avledningssuffix, så kallade képző. Avledningssuffixen förändrar ordens betydelse och kan även göra så att orden byter ordklass, t.ex. *hull* 'falla' *hulladék* 'avfall' och *szép* 'vacker' *szépség* 'skönhet'.

- Pluralis

Pluralis av substantiv uttrycks med pluralmärket "-k" som föregås av bindevokalerna "-a-, -e-, -o- eller -ö-". Pluralmärket är av typen "jel", t.ex. *erdők* 'skogar', *ház-ak* 'hus+PL', *könyv-ek* 'böcker', *virág-ok* 'blommor', *bőr-ök* 'skinn+PL'.

- Possessivsuffix

Possessivsuffixen uttrycker ägare och motsvarar svenskans possessiva pronomen. Suffixen uttrycker både ägarens person och numerus. Possessivsuffixen är av typen "jel". Om det ägda är i singularis fogas följande suffix till ordet:

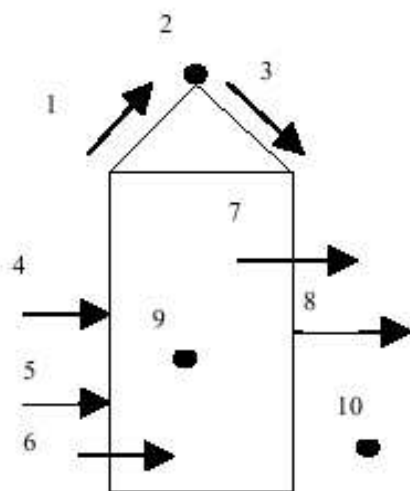
- Första person singularis: -m med bindevokalerna -a-, -e-, -o- eller -ö-
- Andra person singularis: -d med bindevokalerna -a-, -e-, -o- eller -ö-
- Tredje person singularis: -a, -e, -ja eller -je
- Första person pluralis: -nk med bindevokalerna -u- eller -ü-
- Andra person pluralis: -tok med bindevokalerna -a- eller -o-, -tek med bindevokalen -e-, -tök med bindevokalen -ö-
- Tredje person pluralis: -uk, -ük, -juk, jük, -k

Ett exempel är *ház-am* 'mitt hus'. Om det ägda är i pluralis sätts "-i" före possessivsuffixen men efter bindevokalerna, t.ex. *könyv-eid* 'dina böcker'.

- Possessivmärke

Possessivmärket motsvarar svenskans genitiv och fogas till ägaren. De uttrycks med "-é" i singularis och "-éi" i pluralis. Possessivmärket är av typen "jel", t.ex. *férfi-é* 'mannens'.

1. Riktning från, övre position, sublativ, 'upp på'.
2. Befintlighet, övre position, superessiv, 'på'.
3. Riktning från, övre position, delativ, 'ned från'.
4. Riktning mot, yttre position, allativ, 'till'.
5. Riktning mot, yttre position, terminativ, 'fram till, men inte längre'.
6. Riktning mot, inre position, illativ, 'in i'.
7. Riktning från, inre position, elativ, 'från'.
8. Riktning från, yttre position, ablativ, 'ut ur'
9. Befintlighet, inre position, inessiv, 'i'.
10. Befintlighet, yttre position, adessiv, 'vid'.



Figur 2.1: "Illustration av ungerskans lokalkasussystem"

Nedan följer kasussuffixen. De är samtliga av typen "rag". Om inget annat anges används inte bindevokaler framför kasussuffixen.

- Akkusativ

Ungerska substantiv kan uttrycka objektsform (akkusativ), genom att suffixet "-t" läggs till stammen om den slutar på "-j, -ly, -l, -n, -ny, -r, -s, -sz eller -z". Det finns dock undantag, t.ex. *tej-et* 'mjölk+ACC'. Om stammen slutar på någon av de övriga konsonanterna läggs bindevokalerna "-a-" eller "-e-" mellan stammen och suffixet, t.ex. *asztal-t* 'bord+ACC' och *asztal-ok-at* 'bord+PL+ACC'

För att uttrycka befintlighet och rörelse i olika riktningar använder ungerskan lokalkasus. Lokalkasussystemet är tredelat. Det innebär att "riktning mot", "befintlighet" och "riktning från" kan uttryckas. Det finns också tre positioner "inre", "yttre" och "övre", se figur 2.1.

- Illativ

"Riktning mot" och "inre position" kallas illativ. Det motsvarar svenskans "in

i” och uttrycks med suffixet ”-ba/-be”, t.ex. *a ház-ba* ’in i huset’ och *a szem-be* ’in i ögonen’.

- Inessiv
”Befintlighet” och ”inre position” uttrycks med suffixet inessiv som motsvarar svenskans ”i” och uttrycks med ”-ban/-ben”, t.ex. *a szék-ben* ’i stolen’ och *az ablak-ban* ’i fönstret’.
- Elativ
Elativsuffixen ”-ból/-ből” uttrycker ”riktning från” och ”inre”. På svenska skulle ”från” användas, t.ex. *Svédország-ból* ’från Sverige’, *az erdő-ből* ’från skogen’.
- Allativ
För ”riktning mot” och ”yttre position” används allativsuffixet. Allativ motsvarar svenskans ”till” och har varianterna ”-hoz/-hez/-höz”, t.ex. *a ház-hoz* ’till huset’, *a szék-hez* ’till stolen’ och *az erdő-höz* ’till skogen’.
- Terminativ
Även terminativ används för ”riktning mot” och ”yttre position”. Om ett ord har ett terminativsuffix betyder det ”fram till men inte längre”. Suffixet är ”-ig”, t.ex. *a ház-ig* ’fram till huset’.
- Adessiv
Yttre befintlighet markeras med adessivsuffixet. De skulle på svenska skrivas med prepositionen ”vid” och uttrycks på ungerska med ”-nál/-nél”, t.ex. *az ablak-nál* ’vid fönstret’, *a repülőgép-nél* ’vid flygplanet’.
- Ablativ
Ablativsuffixet ”-tól/-től” uttrycker ”riktning från” och ”yttre position”. På svenska skulle man använda prepositionerna ”från” eller ”av”, t.ex. *indulat-tól* ’av vrede’ och *a kert-től* ’från trädgården’.
- Sublativ
Sublativ uttrycker ”riktning från” och ”övre position”. Det motsvarar svenskans preposition ”upp på” och ”till”. Sublativ har suffixet ”-ra/-re”, t.ex. *a tető-re* ’upp på taket’ och *Magyarország-ra* ’till Ungern’.
- Superessiv
”Befintlighet” och ”övre position” uttrycks med suffixet superessiv. På svenska skulle ”på” användas. Superessiv har varianterna ”-n, -on, -en, -ön”. ”-n” används efter vokalfinala stammar. De övriga används efter konsonantfinala stammar. Ord som innehåller främre vokaler har labial vokalharmoni vid superessiv. Därför får ord som har ”ü” eller ”ö” i sista stavelsen suffixet ”-ön”, t.ex. *a fül-ön* ’på örat’, övriga exempel: *a tető-n* ’på taket’, *az asztal-on* ’på bordet’ och *a szék-en* ’på stolen’.
- Delativ
Delativ används för ”riktning från” och ”övre position”. Delativ motsvaras av svenskans ”ned från” eller ”om”. Delativ uttrycks med suffixet ”-ról/-ről”, t.ex. *a ház-ról* ’ned från huset’ och *a tető-ről* ’ned från taket’.

- **Dativ**
Dativ uttrycks med suffixet ”-nak/-nek”. På svenska uttrycks dativ med prepositionen ”åt”, t.ex. *Anná-nak* ’åt Anna’ och *Péter-nek* ’åt Peter’.
- **Essiv-modal**
Essiv modal har suffixet ”-ul, -ül” och används oftast på adjektiv. Vid substantiv används det för att uttrycka en handlings resultat. Essiv modal motsvarar svenskans ”såsom”, t.ex. *a hajó-ul* ’såsom båten’ och *a szék-ül* ’såsom stolen’.
- **Essiv-formalis**
Essiv formalis har suffixet ”-ként” och uttrycker tillstånd. Detta suffix motsvarar ”som”, t.ex. *tanár-ként* ’som lärare’.
- **Translativ-faktiv**
För att uttrycka att något förändras till något annat används translativ-faktiv-suffixet ”-vá/-vé”. På svenska skulle ”till” användas. ”v” i suffixet assimileras till stammens sista bokstav om den är en konsonant, t.ex. *jég-gé* ’till is’ och *tészta-vá* ’till deg’.
- **Instrumental-komitativ**
Instrumental-komitativ uttrycks med suffixet ”-val/-vel”. ”v” i suffixet assimileras till stammens sista bokstav om den är en konsonant. Suffixet betyder ”tillsammans med”, t.ex. *Anná-val* ’med Anna’ och *Péter-rel* ’med Peter’.
- **Instrumental-sociativ**
Instrumental-sociativ har suffixet ”-stul/-stül”. Före ”-stul” sätts bindevokaler ”-a-” eller ”-o-” vid konsonantfinala stammar, före ”-stül” sätts ”-e-” eller ”-ö-”. Sociativ motsvarar ”med”, t.ex. *ruhá-stul* ’med kläderna på’ och *felesége-stiül* ’med hans/hennes hustru’.
- **Kausal-final**
Suffixet ”-ért” motsvarar svenskans ”för...skull”, t.ex. *Péter-ért* ’för Peters skull’.
- **Distributiv**
Distributiv motsvarar ”per” på svenska och har suffixet ”-nként” vilket föregås av bindevokaler ”-o-, -e-, -ö-” vid behov, t.ex. *ember-enként* ’per person’.
- **Temporal**
Temporal uttrycks med ”-kor” och motsvarar ”vid” när det handlar om tid, t.ex. *húsvét-kor* ’vid påsk’.
- **Distributiv-temporalis**
Distributiv-temporalis uttrycks med ”-onta/-anta/-ente” för att säga att något görs regelbundet, t.ex. *nap-onta* ’varje dag’, *nyar-anta* ’varje sommar’ och *het-ente* ’varje vecka’.

3 Automatisk morfologisk segmentering och analys

Detta kapitel presenterar automatisk morfologisk segmentering och analys. Både regelbaserad analys och statistisk analys beskrivs. Dessutom ges exempel på olika datorlingvistiska applikationer där automatisk morfologisk segmentering eller analys används.

Det går att göra morfologisk segmentering och analys manuellt men det är tidskrävande och kostnadsineffektivt och dessutom blir inkonsekventa fel. Därför görs morfologisk segmentering och analys oftast automatiskt. Automatisk morfologisk segmentering och analys används inom olika datorlingvistiska applikationer för att begränsa lexikonstorleken i programmen. Om man inte måste lägga till alla böjningsformer av ett ord i ett lexikon utan endast grundformen blir lexikonet mindre och tar då också upp mindre minne. Dessutom kan okända ord, dvs. ord som inte finns i lexikonet segmenteras och analyseras förutsatt att suffixen är kända. Automatisk morfologisk segmentering och analys är dock inte helt problemfritt. I de flesta språk finns regelbunda böjningsformer och stamförändringar såsom omljud och avljud. Det kan finnas komplicerade avlednings- och sammansättningsregler och dessutom kan samma böjningsform höra till olika morfem.

En regelbaserad automatisk analysator använder ett stamlexikon, ett suffixlexikon och en uppsättning regler för fonologiska och ortografiska regelbundenheter. Det finns olika system för regelbaserad analys.

Item and Arrangement, Item and Process och Word and Paradigm är tre klassiska sätt att göra morfologisk analys. I ”Item and Arrangement”-system klassas både stammar och affix som morfem. Morfemen utgör basenheten, man gör upp listor över morfemen och anger möjliga positioner för varje varje morfem. Varje morfem har en underliggande form och dessa sätts ihop till en sträng eller ett träd. Ord byggs upp av morfem på samma sätt som satser byggs upp av ord (Sproat, 1992). Denna modell är bra för enkel konkateneringsmorfologi (Maxwell, 1998) och fungerar bäst för agglutinerande språk.

I ”Item and Process”-system byggs orden upp av ordbildningsregler (Maxwell, 1998). Det är endast stammarna som ses som morfem. Ord byggs upp genom att man succes-

sivt applicerar ordbildningsregler på stammen (Sproat, 1992). Suffixen är inte morfologiska enheter utan tillhör de regler i vilka de presenteras. Stammen kan ha en underliggande form men alla andra morfem motsvarar processer som påverkar den fonologiska formen. En ordbildningsregel kan ses som en relation där stammen är input, stammens fonologiska form ändras, morfosyntaktiska egenskaper läggs till, outputn blir en böjningsform av ordet.

I "Word and Paradigm"-system ställs böjningsparadigm för orden upp. Ord bildas genom att en regel appliceras på ett redan existerande ord. Varje lexem associeras med vissa egenskaper. Man utgår från en stam och plockar böjningsformerna ur ett paradigm. "Word and Paradigm" separerar ordformer från morfosyntaktiska ord. Morfosyntaktiska ord byggs upp utan uttal. Ordformer byggs upp utan betydelse. De två formerna sammanfogas med hjälp av paradigm (Sproat, 1992). Denna typ av system fungerar bra för flekterande språk. Samtliga tre ovanstående system presenterades först av Hockett (1954). Word and Paradigm har beskrivits ytterligare av Robins (1959).

En annan typ av regelbaserad analys är Kimmo Koskenniemis tvånivåmorfologi (Koskenniemi, 1983). Tvånivåmorfologisystemen består av lexikon i två nivåer, en ortografisk nivå och en lexikal nivå. De två nivåerna kopplas ihop av transduktorer som möjliggör både analys och generering. Förutom lexikonet krävs tvånivåregler där en viss ortografisk symbol realiserar som en viss lexikal symbol i en bestämd höger- och vänsterkontext, t.ex. "a" realiserar som "b" om vänsterkontexten är "c" och högerkontexten är "d":

a:b <=> c_d.

Automatisk morfologisk segmentering och analys kan också göras av system som använder statistik. Statistiska system ger oftast en helt disambiguerad output men det finns oftare fler fel i outputn än vad det gör i motsvarande regelbaserade systems output (Karlsson och Karttunen, 1996). Den vanligaste varianten är probabalistiska (stokastiska) metoder för morfologisk disambiguering. Statistiska system innehåller ett frekvenslexikon. I frekvenslexikonet finns frekvensinformation för n-gram som används för att hitta suffix (Kornai, 1992). Probabalistiska system kan få sin frekvensinformation från en handtaggad träningskorpus. Man kan också använda dolda Markovmodeller för att träna systemen på en otaggad korpus.

Vissa program för automatisk morfologisk analys kan också användas för att generera ord. Det görs då givet ett lemma, böjningsaffix, avledningssuffix och morfotaxregler.

Automatisk morfologisk analys kan användas inom ordklasstagning. I ett ordklasstagningssystem kan information om vilka suffix som hör till vilken ordklass läggas till. Ett ord kan då inte taggas som verb om det har ett suffix som tillhör substantiven. I många ordklasstagningssystem ingår morfologisk taggning som en del av systemet. Man får alltså ut både ordklasstaggar och morfologiska taggar.

Det finns många olika datorlingvistiska applikationer som använder automatisk morfologisk segmentering eller analys i någon mån. Ett av de enklaste användningsområdena är trunkering. Vid trunkering delas ett ord upp i en stamdel och en suffixdel men suffixen segmenteras inte. Detta kan göras för att hitta ordstammar och är användbart inom informationssökning. I ett informationssökningssystem trunkeras sökordet så att endast stammen blir kvar. Det leder till att fler träffar kan fås vid en sökning på ordet eftersom ordets olika böjningsformer kan hittas.

Inom ordbehandling används automatisk segmentering för att skapa effektiva rättstavningsprogram. För att göra detta används ett lexikon som består av rotmorfem, en uppsättning affix och morfotaxregler. Även automatisk avstavning kan använda automatisk morfologisk segmentering. Genom att segmentera ett ord till dess morfer hamnar avstavningen på rätt plats.

Inom maskinöversättning används automatisk morfologisk analys för att minska storleken på de elektroniska lexikonerna. Genom att endast ha stammar i ett lexikon och lista alla möjliga suffix blir lexikonet mindre än om man skulle ha samtliga böjningsformer av alla ord i lexikonet.

I text-till-talsystem används morfologisk information för att ge bättre uttal. Två ord som ortografiskt ser likadana ut kan få olika uttal beroende på vilken morfologisk analys det får. Ett exempel är *tomten*, om ordet segmenteras i *tomte-n* får det ett uttal, om det däremot segmenteras i *tomt-en* får det ett annat uttal. Vid taligenkänning blir lexikonet mindre och träningstiden förkortas om man använder ett morfbaserat igenkänningssystem (Trost, 2003).

För agglutinerande språk som ungerskan är automatisk morfologisk analys mycket användbart eftersom många suffix leder till att många kombinationer kan skapas (Váradi och Oravecz, 1999). Den morfologiska analysen är alltså extra viktig för formrika språk. Det finns flera system som använder automatisk morfologisk analys för att analysera ungerska ord. Några av dem beskrivs nedan.

3.1 PC-KIMMO

PC-KIMMO är en morfologisk parser som lanserades 1995. Den är en utveckling från tidigare versioner som alla utvecklats från Kimmo Koskenniemis modell för två-nivå-morfologi (Koskenniemi, 1983). PC-KIMMO består av ett lexikon och regler. I lexikonet finns alla morfem; både stammar och affix i sin lexikala form med specificerade morfotaktiska restriktioner, se exempel 10-12. Två-nivå-reglerna tar hand om fonologiska eller ortografiska variationer, se kapitel 3. I regelfilen finns också alfabetet specificerat. Tillsammans tokeniserar reglerna och lexikonet de ingående orden. Informationen från lexikonet och reglerna unifieras till ett Feature Structure

Tree (Antworth, 1994). PC-KIMMO kan både generera och analysera. Generatoren tar den lexikala formen av ordet som input, applicerar regler och tar fram ytordet. Analysatorn tar ytordet som input, applicerar reglerna, konsulterar lexikonet och får den lexikala formen som output. PC-KIMMO finns att ladda ner som körbart program¹. Man kan själv lägga till sina egna regel- och lexikonfiler. Det finns också exempel på beskrivningar för en rad språk, bland annat finska och turkiska (Antworth, 1990).

3.2 Humor

Humor står för High-speed Unification Morphology. Systemet utvecklades av Gábor Prósztékly med flera (Prósztékly, 1994) och är implementerat i C. Humor utvecklades för ungerska men fungerar idag även för många andra språk (Prósztékly och Balázs, 1999).

Humor är en fullständig morfologisk analysator och generator. Alla morfemkombinationer som skapas körs genom Analysatorn. Lemmatiseraren är en förenklad version av Analysatorn. Den producerar alla möjliga lexikala stammar av en ordform. Humor använder sig av gissningsstrategier. En gissare är en variant av morfologisk analysator som innehåller alla fonologiskt möjliga stammar. Humors gissare har information om ortografiska, morfologiska, morfofonologiska och lexikala egenskaper hos orden. Humor är feltolerant (Prósztékly m.fl., 1994). Om användaren skriver in ett felstavat ord autokorrigeras det till standardortografi innan det analyseras.

Humor tillämpas i en rad språkteknologiska verktyg: en- och tvåspråkiga ordböcker, stavningskontroller, lemmatiserare med mera. De två största kommersiella programmen som skapats med hjälp av Humor-analysatorn är Helyette, en tesaurus som hittar ordstammen, sparar suffixen, letar upp synonymer till aktuell stam och fogar suffixen till synonymstammarna. Den andra är Morphologic's Bi-lingual Dictionary (MoBi-Dic), som översätter mellan engelska och ungerska. Ordboken täcker hela Concise Explanatory Dictionary of Hungarian Language.

3.3 Xerox morfologisk analys, tokenisering och disambiguering

Xerox Research Center Europe har utvecklat morfologiska analysatorer för ett flertal språk (Beesley och Karttunen, 2003). De olika applikationerna är skrivna i Xerox finite-state software (XFST), Xerox egna interaktiva gränssnitt. XFST är ett gränssnitt som ser till att kompilatorer för reguljära uttryck direkt kan få tillgång till Xerox Finite State Calculus. Finite State Calculus är den algoritm som använts för utvecklandet av de olika applikationerna. Dessutom används Lexicon Compiler (LEXC) som är ett deklarativt högnivåspråk för att specificera lexikon och beskriva morfotaxen i språket. XFST-gränssnittet och LEXC-kompilatorn kan användas som verktyg

¹<http://www.sil.org/pckimmo/>

ör att skapa nätverk av finita tillstånd (finite-state network). För att bygga en morfologisk analysator med XFST och Finite State Calculus skapas ett lexikon, morfotaktiska regler, ett morfotaktiskt filter samt regler för ortografiska och fonologiska variationer. Dessa delar implementeras som ett nätverk av finita tillstånd. De regler som bestämmer morfemens form, dvs. regler för ortografiska och fonologiska variationer (alternations) implementeras som finita transduktorer. De kombineras i lexikala transduktorer "lexical transducers". En lexikal transduktor gör både morfologisk analys och morfologisk generering. Det ger effektiva och robusta program som tar upp ett minimum av minne. För fler tekniska detaljer hänvisas intresserade läsare till Beesley och Karttunen (2003).

De två främsta applikationerna som Xerox utvecklat är tokenisering och uppslagning av ord och deras morfologiska analys. Tokeniseraren använder ett nätverk av finita tillstånd för att tokenisera en löpande text i tokens. Uppslagning innebär en morfologisk analys av redan tokeniserad text. Vid uppslagning kan en gissare som innehåller alla fonologiskt möjliga stammar användas. Tokenisering och uppslagning används ofta tillsammans. En fil tokeniseras först av programmet. Uppslagningsmekanismen använder sedan den tokeniserade filen som indata. Utdata blir ordet, grundformen och taggarna. Den morfologiska analysen är grunden i de flesta andra applikationerna.

Xerox har utvecklat kommersiella system för bearbetning av naturliga språk, ordklasstagning, parsning och översättning. Xerox implementationer har visat sig användbara för att bygga storskaliga morfologiska analysatorer och generatorer för agglutinerande språk som t.ex. ungerska. Analysatorn för ungerska skrevs av Agnes Sandor².

²Demo finns på <http://www.xrce.xerox.com/competencies/content-analysis/demos/hungarian>

4 Automatisk morfologisk segmentering och analys av ungerska substantiv

I detta kapitel finns en beskrivning av de korpusar som använts vid skapandet av ett program för automatisk morfologisk segmentering och analys av ungerska substantiv samt en utförlig genomgång av programmets utformning.

4.1 Data

De ungerska korpusar jag använt i mitt arbete är Szeged Corpus (Zoltán m.fl., 2002) och Hungarian National Corpus (Váradi, 1998). Szeged Corpus skapades mellan åren 2000 och 2002 i ramverket Info-communication Technologies and Applications (IKTA) på Szegeduniversitetet. Arbetet hade titeln ”Development of a Part-of-Speech Tagging Method for Hungarian by using Machine Learning Algorithms”. Szegedkorpusen är en morfosyntaktiskt och ordklassannoterad databas. Korpusen består av en miljon ordningångar. Texterna som utgör korpusen kommer från olika genrer: skönlitteratur, tidningsartiklar, texter relaterade till datavetenskap, lagtext samt korta uppsatser skrivna av elever i åldrarna fjorton till sexton år. Ordningångarna bearbetades morfosyntaktiskt med hjälp av Humor morpho-syntactic analyser som utvecklats av MorphoLogic Ltd, se kapitel 3.2 för vidare beskrivning. Då genererades möjliga morfosyntaktiska taggar för varje ord. Sedan ordklasstaggades hela korpusen manuellt och korrekt morfosyntaktisk etikett valdes ut utifrån kontexten. Det är gratis att använda Szegedkorpusen i utbildnings- och forskningssyfte¹.

Hungarian National Corpus (HNC) skapades 1998 på the Department of Corpus Linguistics of the Research Institute for Linguistics of the Hungarian Academy of Sciences. Korpusen består av 153.7 miljoner ord och är uppdelad i fem delkorpusar. Delkorpusarna består av texter från olika specifika genrer. De olika delarna är texter från media, litterära texter, vetenskapliga texter, officiella dokument samt informella texter. HNC är tillgänglig för alla².

4.2 Programmets utformning

PC-KIMMO har valts som skalprogram för att skapa en automatisk morfologisk analysator för ungerska substantiv eftersom man på ett enkelt sätt kan skapa ett program

¹Szegedkorpusen hittas på följande adress: <http://www.inf.u-szeged.hu/projectdirs/hlt/corpus1-en.htm>

²HNC hittas på följande adress: <http://corpus.nyud.hu/mnsz/>

för ett specifikt språk. Den kod som krävs för att programmet överhuvudtaget ska fungera finns redan. Det användaren gör är att lägga till språkspecifik information.

För att göra automatisk morfologisk analys med PC-KIMMO krävs följande:

- en ordlista med ord i deras grundform
- en lista med samtliga suffix
- ett lexikon som beskriver morfotaxen
- en regelbil som tar hand om fonologiska variationer i språket

För att skapa en automatisk morfologisk analysator för ungerska substantiv samlades en ordlista med substantiv i deras grundform. Till det användes Szegedkorpussen. Den består av ett antal filer med ord i grundform och deras ordklasstag. Samtliga filer sparades och de rader med substantivtagg valdes ut. Szegedkorpussen innehåller drygt 17000 unika substantiv i grundform. Samtliga lades i en fil "Substantivfilen".

För att PC-KIMMO ska kunna hantera orden krävs information om vad som är det lexikala ordet, vilket sublexikon ordet tillhör, vad som får följa efter ordet och den information som ska visas efter analys, se exempel 10. "Lexical item" är det lexikala ordet. I lexikonfilen finns två sublexikon "Noun" och "Suffix". "Suffix" har olika dellexikon där de olika suffixtyperna definieras. I fältet "Sublexicon" anges alltså vilket lexikon ordet tillhör. I fältet "Alternation" anges vad som får följa efter ordet, om ordet är en stam får alla suffix följa efter ordet, se exempel 10, är ordet däremot ett kasussuffix får ingenting följa efter ordet, se exempel 11. "Gloss" anger vad som ska skrivas ut efter analys. För stammar skrivs själva ordet ut, för suffix skrivs den morfologiska taggen ut.

Exempel 10 Substantivingång i PC-KIMMOs lexikon

```
\lf asztal %lexical item
\lx NOUN %sublexicon
\alt Suffix %alternation
\gl asztal %gloss
```

En fil med samtliga suffix som de ungerska substantiven kan ha skapades manuellt. Varje suffix lades in tillsammans med de varianter av suffixen som finns, se uppräkningsen av suffix i kapitel 2.1.2. För att PC-KIMMO ska kunna hantera suffixen lades förutom själva suffixet information in om vilken typ av suffix som är aktuellt, om ytterligare suffix får följa efter samt förkortningen av suffixnamnet, se exempel 11. De flesta suffix fick mer än en ingång eftersom varje variant (olika bindevokaler) av suffixet kräver en egen ingång, se exempel 12. De suffixtaggar som valdes utgår från de taggar som finns i "Hungarian National Corpus" eftersom gemensamma taggar underlättar utvärderingsarbetet. Det blir också lättare för en användare som är van att arbeta med "Hungarian National Corpus" att använda programmet. "Hungarian National Corpus" är den största ungerska korpussen som finns idag. Totalt innehåller

suffixlexikonet 186 ingångar.

Exempel 11 Suffixingång i PC-KIMMOs lexikon

```
\lf +ban
\lx INESSIVE
\alt End
\gl +INE
```

Exempel 12 Samtliga ingångar för pluralsuffixet

```
% Plural lexicon
\lf +k
\lx PLURAL
\alt Rag
\gl +PL

\lf +ak
\lx PLURAL
\alt Rag
\gl +PL

\lf +ek
\lx PLURAL
\alt Rag
\gl +PL

\lf +ok
\lx PLURAL
\alt Rag
\gl +PL

\lf +ök
\lx PLURAL
\alt Rag
\gl +PL
```

Lexikonet innehåller så kallade ”alternations” som talar om hur stammarna och suffixen får kombineras med varandra. Det är alltså lexikonet som beskriver den morfotaktiska analysen. I lexikonet definieras samtliga suffix. En indelning av vilka suffix som är ”rag” och vilka som är ”jel” görs eftersom det underlättar när programmet ska avgöra vilka suffix som får kombineras och i vilken ordning de får stå i ordet, se kapitel 2.1.2.

Den teoretiska bakgrunden har varit grund för de regler som skapats. Regelfilen innehåller det ungerska alfabetet, se exempel 1, dvs. en definition av vilka bokstäver som får finnas med i orden som ska analyseras. I regelfilen finns också regler som

tar hand om fonologiska variationer. Ord som slutar på "a" eller "e" får i ungerskan förlängd slutvokal då ett suffix fogas till ordet, se exempel 4. En regel för att återställa de förlängda vokalerna när suffixen plockas bort har skapats. Detta görs för att PC-KIMMO ska kunna känna igen ordstammarna. Om ordet som ska analyseras t.ex. är *macskával* 'med katten', så består det av *macska* med förlängd slutvokal och *val* som är inessivsuffixet. Regeln gör om *macská* till *macska* 'katt' när suffixet plockats bort, se exempel 13.

Exempel 13 Ord med förlängd slutvokal före suffix

macskával	'med katt'
macská-INS	<i>macska</i> med förlängd slutvokal +INS
macská => macska	'katt'

För att köra programmet laddas regler och lexikon in i PC-KIMMO. Sedan kan man antingen analysera ett enskilt ord eller en hel fil. Resultet skrivs ut på skärmen eller i en fil beroende på vilka kommandoflaggor som används. Ett körningsexempel visas i exempel 14. I exemplet står "r" för recognize vilket betyder att analys ska göras. *asztalon* betyder 'på bordet' och är ordet som ska analyseras. I resultatet visas först grundformen + suffixet och sedan kommer grundformen + suffixförkortningen

Exempel 14 Testkörning

PC-KIMMO>r asztalon
asztal+on [asztal+SUP]

5 Utvärdering

I detta kapitel presenteras utvärderingsmetoden och resultatet av utvärderingen av programmet.

5.1 Utvärderingsmetod

Utvärderingen gjordes för att ta reda på hur stor del av en testsvit som segmenteras och analyseras korrekt, det vill säga programmets precision mättes.

En testsvit skapades genom att 500 substantiv plockades ut slumpvist och automatiskt från "Hungarian National Corpus", se kapitel 4.1 för mer information om korpusen. För testsviten valdes en annan korpus än den som ligger till grund för substantivlexikonet i programmet för att utvärderingen skulle bli så självständig som möjligt. "Hungarian National Corpus" finns tillgänglig på internet. Ur korpusen kan man plocka ut som mest 500 ord åt gången. Det görs genom att man specificerar en sökning genom att välja vilka ordklasser och morfologiska former som ska sökas. Dessutom får man ange hur många ord som ska hittas och ur vilka subkorporar orden ska plockas, se figur 5.1. I testsviten finns de korrekta morfologiska taggarna med, vilket underlättar arbetet med att undersöka resultatet av testkörningen.

Programmet kördes på testsviten. Resultatet gicks igenom manuellt och samtliga ord lades i någon av kategorierna: korrekt taggade ord med en analys, korrekt taggade ord med flera analyser, felaktigt taggade ord och ej taggade ord.

5.2 Resultat av utvärderingen

I tabell 1 redovisas resultatet av testkörningen.

Tabell 1 resultat av testkörning

1. word-form part-of-speech: nominal ... MSD-code:

part-of-speech: type: number: possessive: > poss. number: > pluralizer: anaphoric: > pluralizer: case:

noun ANY ANY ANY ANY ANY ANY ANY

only:

A random sample of 500 items with 1 word context.

Besides the word-form only on target word the stem and the MSD-code will appear. Attributes in small window.

Sorting: in original order. Bibliography: in small window.

Subcorpus: Distribution by subcorpora. Author:

Press
Literature
Science
Official
Personal

start search

Figur 5.1: sökning i "Hungarian National Corpus"

	antal ord i %	antal ord
Korrekt taggade ord	73,6%	368
en tagg	62,6%	313
flera taggar	11%	55
Felanalyserade ord	0,2%	1
Ej taggade ord	26,2%	131

313 av 500 ord fick den korrekta analysen, vilket motsvarar 62,6%. 55 ord (11%) fick två eller tre analyser varav en av dessa analyser var den korrekta. Om man lägger ihop dessa resultat får man ett mått på hur väl programmet fungerar. För denna testsvit blev det totala antalet korrekta analyser 73,6%.

Bland de ord som fått mer än en analys utmärker sig de där valet står mellan att ordet ska få analysen possessiv, tredje person singular, det ägda i singular (PSe3) eller possessivmärke (POS). Hela 29% (16 st) av orden som fått mer än en analys var sådana. I samtliga fall var PSe3 den korrekta lösningen. Anledningen till att problemet uppstår är att "é" i de aktuella orden skulle kunna markera både PSe3 och POS eftersom bindevokals-e förlängs innan suffix, se exempel 4. De är alltså tvetydiga. POS är dock en ovanligare form. I exempel 15 visas resultatet för ett sådant ord.

Programmet kan i 23,6% (13 ord) av fallen inte skilja mellan "a" och "e" som markerar possessiv, tredje person singular, det ägda i singular (PSe3) och bindevokalerna "a" och "e" framför akkusativ-t, se exempel 16. Det går inte att uttala sig utifrån detta testmaterial om vilken av formerna som är vanligast.

Något som också ställer till problem är att både plural (PL) och possessiv, tredje person plural, det ägda i singular (PSt3) kan markeras med "k", se exempel 17. Denna tvetydighet omfattar 21,8% (12 st) av orden som fått mer än en analys.

Bland de återstående 25,6% (14 st) av orden går det inte att hitta en gemensam anledning till att de fått mer än en analys. De listas nedan, den korrekta analysen är markerad med "R" framför ordet:

- értékben
ér+tek+ben [ér+PSt2+INE]
R érték+ben [érték+INE]
- székhelyen
székhely+e+n [székhely+PSe3+SUP]

- R székhely+en [székhely+SUP]
- bővítését R bővítés+e+t [bővítés+PSe3+ACC]
bővítés+et [bővítés+ACC]
bővítés+é+t [bővítés+POS+ACC]
- kárt
R kár+t [kár+ACC]
kar+t [kar+ACC]
- törvényen
törvény+e+n [törvény+PSe3+SUP]
R törvény+en [törvény+SUP]
- helyszínen
R helyszín+e+n [helyszín+PSe3+SUP]
helyszín+en [helyszín+SUP]
helyszín+é+n [helyszín+POS+SUP]
- élet él+e+t [él+PSe3+ACC]
él+et [él+ACC]
R élet [élet+NOM]
- ágya
R ágy+a [ágy+PSe3]
agy+a [agy+PSe3]
- házakat
R ház+ak+at [ház+PL+ACC]
haza+k+at [haza+PL+ACC]
haza+k+at [haza+PSt3+ACC]
- növekedését
R növekedés+e+t [növekedés+PSe3+ACC]
növekedés+et [növekedés+ACC]
növekedés+é+t [növekedés+POS+ACC]
- sírjánál
R sír+ja+nál [sír+PSe3+ADE]
sír+ja+nal [sír+PSe3+INS]
- bevezetését
R bevezetés+e+t [bevezetés+PSe3+ACC]
bevezetés+et [bevezetés+ACC]
bevezetés+é+t [bevezetés+POS+ACC]
- érték
ér+tek [ér+PSt2]
R érték [érték+NOM]
- megértését
R megértés+e+t [megértés+PSe3+ACC]
megértés+et [megértés+ACC]
megértés+é+t [megértés+POS+ACC]

Exempel 15-17 visar på ord i testsviten som fått mer än en analys

Exempel 15

fényében		
fény+e+ben	[fény+PSe3+INE]	'i hans/hennes glans'
fény+é+ben	[fény+POS+INE]	felaktig analys

Exempel 16

körzetet		
körzet+et	[körzet+ACC]	'omfattning+ACC'
körzet+e+t	[körzet+PSe3+ACC]	felaktig analys

Exempel 17

listák		
lista+k	[lista+PL]	'listor'
lista+k	[lista+PSt3]	felaktig analys

5.2.1 Felanalyserade ord

0,02% (1 ord) har fått en felaktig analys. Ordet är művészeit som fått analysen művész+ei+t [művész+PSt1i+ACC] 'våra konstnärer+ACC'. Ordet borde ha fått analysen művészeit [művész+PSe3i+ACC] 'hans/hennes konstnär+ACC'.

5.2.2 Ej analyserade ord

26,2% (131) av orden i testkorpussen har inte fått någon analys överhuvudtaget. Det beror i de flesta fall på att grundformen inte finns med i programmets substantivlista, 109 av orden som inte fått någon analys har inte sin grundform i substantivlexikonet. Egennamn (personnamn och namn på geografiska platser) utmärker sig bland dessa ord. 38 egennamn återfinns bland orden som inte fått någon analys. Av de övriga 21 orden är 17 ord oregelbundet böjda, tre ord är utländska och ett ord är ett nonsensord.

6 Diskussion

Programmet har utvärderats. Resultatet visar att 73,6% av orden i en slumpmässigt utvald testsvit på femhundra ord får en korrekt analys. De ord som inte fått en korrekt analys har gått igenom manuellt. Bland orden som inte fått en korrekt analys återfinns ord som inte fått någon analys alls och ord som fått en felaktig analys. Det är allvarligare om ett ord får en felaktig analys än ingen analys alls. Har ett ord ingen analys är det lättare att upptäcka och korrigera manuellt för användaren. Det är endast ett ord i den aktuella testsviten som fått en felaktig analys.

En stor del av orden som inte fått någon analys har inte fått det på grund av att stammen inte finns med i substantivlexikonet. Genom att göra fler testkörningar på ord från "Hungarian National Corpus" kan man fånga upp dessa och lägga till dem i lexikonet och på så vis få ett bättre resultat vid framtida körningar. Får man tillgång till ytterligare ungerska korpusar kan man lägga till även ord ur dem. Att lägga till fler ord i substantivlexikonet är troligen det bästa sättet att öka programmets precision.

De ord som inte fått någon analys trots att stammen finns i substantivlexikonet är ord som böjs oregelbundet. Att lägga in vanliga oregelbundna substantiv direkt i programmet skulle öka programmets förmåga att analysera ord. Det kräver dock en lista över vanliga oregelbundna ord.

Bland de ord som fått mer än en analys utmärker sig vissa kombinationer av analyser. Bland annat har många ord fått analysen possessiv tredje person singular, det ägda i singular (PSe3) och possessivmärke, det ägda i singular (POS), se exempel 15. I det fallet var den korrekta analysen PSe3 för samtliga ord i testsviten. Man skulle kunna koppla statistik till programmet och på så vis alltid plocka ut PSe3 som korrekt analys när valet står mellan PSe3 och POS. Detta går även att applicera på andra liknande par av analyser där den ena analysen är mycket mer frekvent än den andra.

I framtiden kan man fortsätta arbeta med att lägga till övriga ordklasser för att få ett heltäckande och användbart program. Programmet måste då utökas med stamlexikon för de olika ordklasserna, listor med nya suffixtyper samt ytterligare regler för fonologiska oregelbundenheter.

Man kan använda PC-KIMMO för generering. För att programmet ska kunna genere-

ra ord krävs en utbyggnad av reglerna. Det krävs regler som tar hand om vokalharmonin. Man måste specificera vilka ord som får kombineras med vilka suffixvarianter. *Asztal-on*, *asztal-SUP*, 'på bordet' ska kunna genereras utan att man samtidigt får de felaktiga formerna *asztal-n*, *asztal-en* och *asztal-ön*.

7 Sammanfattning

I denna uppsats behandlas automatisk morfologisk segmentering och analys av ungerska substantiv. Ungerska substantivs morfologi har presenterats. Automatisk morfologisk segmentering och analys har beskrivits. Dessutom har ett program skapats som analyserar ungerska substantiv. Programmet skapades genom att lexikon för substantiv och suffix samt regler lades till i PC-KIMMO. Programmet har också utvärderats. En testkörning på 500 slumpmässigt utvalda ord ur "Hungarian National Corpus" visar att 73,6% av orden får en korrekt analys.

Litteraturförteckning

- Antworth, Evan L. Pc-kimmo: a two-level processor for morphological analysis, 1990. Dallas.
- Antworth, Evan L. Morphological parsing with a unification-based word grammar. University of Texas at Arlington, 1994.
- Beesley, Kenneth och Karttunen, Lauri. *Finite State Morphology*. CSLI Publications, 2003.
- Dugántsy, Mária. *Ungersk Grammatik 1*. Universitetsstryckeriet, 1997.
- Hockett, Charles. Two models of grammatical description. *Word* 10, 1954.
- Karlsson, Fred och Karttunen, Lauri. *Survey of the State of the 4Art in Human Language Technology*, kapitel 3. Center for Spoken Language Understanding, Oregon Graduate Institute, USA, 1996.
- Kornai, András. Frequency in morphology. *Approaches to Hungarian*, 4:246–268, 1992.
- Koskenniemi, Kimmo. Two-level morphology: a general computational model for word-form recognition and production. I: *Publication No. 11*. University of Helsinki: Department of General Linguistics., 1983.
- Lavotha, Odön och Lavotha, Csilla. *Ungersk grammatik*. Almqvist & Wiksell, 1973.
- Maxwell, Mike. Two theories of morphology, one implementation. Summer Institute of Linguistics, Inc, 1998.
- Megyesi, Beáta. A short descriptive grammar for hungarian. unpublished, 1998.
- Prószéky, Gábor. Industrial applications of unification morphology. I: *Proceedings of the 4th Conference on Applied Natural Language Processing*, ss 157–159, 1994.
- Prószéky, Gábor och Balázs, Kis. A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. I: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic*, ss 261–268, 1999.
- Prószéky, Gábor, Miklós, Pál, och Tihanyi, László. Humor-based applications. I: *Proceedings of COLING-94*, ss 1241–1244, 1994.
- Robins, Robert. In defence of wp. *Transactions of the Philological Society*, 1959.
- Sproat, Richard. *Morphology and Computation*. MIT Press, 1992.

- Storlind, Eugen. *Ungersk grammatik*. Storlinds förlag, 2002.
- Trost, Harald. *The Oxford Handbook of Computational Linguistics*, kapitel 1, ss 25–47. Oxford University Press, Oxford, 2003.
- Váradi, Tamás. Hungarian national corpus. Department of Corpus Linguistics of the Research Institute for Linguistics of the Hungarian Academy of Sciences, 1998.
- Váradi, Tamás och Oravecz, Csaba. Morphosyntactic ambiguity and tagset design for hungarian. Teknisk rapport, Linguistics Institute Hungarian Academy of Sciences, 1999.
- Zoltán, Alexin, Csirik, János, Gyimóthy, Tibor, och Prózéky, Gábor. Szeged corpus 1.0. University of Szeged, Department of Informatics, 2002.

A Kod

Regelfilen

COMMENT %

% the Alphabet consists of all Hungarian letters, "-" which is the hyphen symbol and "+" which marks morpheme boundary

ALPHABET A a Á á B b C c Cs cs D d Dz dz Dzs dzs E e É é F f G g Gy gy H h I i Í í J j K k L l Ly ly M m N n Ny ny O o Ó ó Ö ö Õ õ P p Q q R r S s Sz sz T t Ty ty U u Ú ú Ü ü Ű ű V v W w X x Y y Z z Zs zs - +

% the Null symbol is used for deletions and insertions.

NULL 0

% @ is a "wildcard" symbol

ANY @

% # marks word boundary

BOUNDARY #

% Cons defines which letters are consonants

SUBSET Cons B b C c Cs cs D d Dz dz Dzs dzs F f G g Gy gy H h J j K k L l Ly ly M m N n Ny ny P p Q q R r S s Sz sz T t Ty ty V v W w X x Z z Zs zs

% Vow defines which letters are vowels

SUBSET Vow a á e é i í o ó ö õ u ú ü û y

% the below rules show how the letters in the alphabet should be represented. The rule is divided into three to make it readable

RULE "default" 1 41

A a Á á B b C c Cs cs D d Dz dz Dzs dzs E e É é F f G g Gy gy H h I i Í í J j K k L l Ly ly @

```

A a Á á B b C c Cs cs D d Dz dz Dzs dzs E e É é F f G g Gy gy
H h I i Í í J j K k L l Ly ly @
1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

```

```

RULE "default" 1 37
M m N n Ny ny O o Ó ó Ö ö Õ õ P p Q q R r S s Sz sz T t Ty ty U
u Ú ú Ü ü Ũ û @
M m N n Ny ny O o Ó ó Ö ö Õ õ P p Q q R r S s Sz sz T t Ty ty U
u Ú ú Ü ü Ũ û @
1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

```

```

RULE "default" 1 15
V v W w X x Y y Z z Zs zs - + @
V v W w X x Y y Z z Zs zs - 0 @
1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

```

%The rules changes long vowels to short where they have been lenghtened before suffixes

```

RULE "a:á <=>_[Cons#|Cons Vow#|Cons Vow Cons#]" 1 88
a B b C c Cs cs D d Dz dz Dzs dzs F f G g Gy gy H h J j K k L l
Ly ly M m N n Ny ny P p Q q R r S s Sz sz T t Ty ty V v W w X x Z z Zs zs
A a E e É é I i Í í O o Ó ó Ö ö Õ õ U u Ú ú Ü ü Ũ û Y y @
á B b C c Cs cs D d Dz dz Dzs dzs F f G g Gy gy H h J j K k L l
Ly ly M m N n Ny ny P p Q q R r S s Sz sz T t Ty ty V v W w X x Z z Zs zs
A a E e É é I i Í í O o Ó ó Ö ö Õ õ U u Ú ú Ü ü Ũ û Y y @
1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

```

```

RULE "e:é <=>_[Cons#|Cons Vow#|Cons Vow Cons#]" 1 84
e B b C c Cs cs D d Dz dz Dzs dzs F f G g Gy gy H h J j K k L l
Ly ly M m N n Ny ny P p Q q R r S s Sz sz T t Ty ty V v W w X x Z z Zs zs
A a Á á E e I i Í í O o Ó ó Ö ö Õ õ U u Ú ú Ü ü Ũ û Y y @
é B b C c Cs cs D d Dz dz Dzs dzs F f G g Gy gy H h J j K k L l
Ly ly M m N n Ny ny P p Q q R r S s Sz sz T t Ty ty V v W w X x Z z Zs zs
A a Á á E e I i Í í O o Ó ó Ö ö Õ õ U u Ú ú Ü ü Ũ û Y y @
1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

```

END

Lexikonfilen

```

% the alternations describe the morphotactic analysis
% suffix is a list of all noun suffixes

```

ALTERNATION Initial INITIAL
 ALTERNATION Noun NOUN
 ALTERNATION Suffix NOMINATIVE PLURAL POSSESSIVE ACCUSATIVE ILLATIVE INESSIVE
 ELATIVE ALLATIVE TERMINATIVE ADESSIVE ABLATIVE SUBLATIVE SUPERESSIVE DELATIVE
 DATIVE ESSIVE FORMAL FACTIVE INSTRUMENTAL SOCIATIVE CAUSAL DISTRIBUTIVE
 TEMPORAL DISTEMP POSSMARK

% ''rag'' is a list of all ''rag'' suffixes i.e. the suffixes that must
 be last in the word

ALTERNATION Nominative End
 ALTERNATION Plural RAG
 ALTERNATION Possessive RAG
 ALTERNATION Possmark RAG
 ALTERNATION Rag ACCUSATIVE ILLATIVE INESSIVE ELATIVE ALLATIVE
 TERMINATIVE ADESSIVE ABLATIVE SUBLATIVE SUPERESSIVE DELATIVE DATIVE
 ESSIVE FORMAL FACTIVE INSTRUMENTAL SOCIATIVE CAUSAL DISTRIBUTIVE
 TEMPORAL DISTEMP End

ALTERNATION Accusative End
 ALTERNATION Illative End
 ALTERNATION Inessive End
 ALTERNATION Elative End
 ALTERNATION Allative End
 ALTERNATION Terminative End
 ALTERNATION Adessive End
 ALTERNATION Ablative End
 ALTERNATION Sublative End
 ALTERNATION Superessive End
 ALTERNATION Delative End
 ALTERNATION Dative End
 ALTERNATION Essive End
 ALTERNATION Formal End
 ALTERNATION Factive End
 ALTERNATION Instrumental End
 ALTERNATION Sociative End
 ALTERNATION Causal End
 ALTERNATION Distributive End
 ALTERNATION Temporal End
 ALTERNATION Distemp End
 ALTERNATION End End

% the FIELDCODE declarations define which code every type of field
 should have in the lexical entry.

FIELDCODE lf U %lexical item
 FIELDCODE lx L %sublexicon
 FIELDCODE alt A %alternation

FIELD CODE gl G %gloss

% file of nouns

INCLUDE nouns.lex

%file of suffixes

INCLUDE suffixes.lex

END

Suffixlexikonet

Nominative lexicon

\lf 0

\lx NOMINATIVE

\alt End

\gl +NOM

% Plural lexicon

\lf +k	\lf +ak	\lf +ek	\lf +ok	\lf +ök
\lx PLURAL	\lx PLURAL	\lx PLURAL	\lx PLURAL	\lx PLURAL
\alt Rag	\alt Rag	\alt Rag	\alt Rag	\alt Rag
\gl +PL	\gl +PL	\gl +PL	\gl +PL	\gl +PL

% Possessive lexicon

\lf +m	\lf +am	\lf +em	\lf +om	\lf +öm
\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE
\alt Rag	\alt Rag	\alt Rag	\alt Rag	\alt Rag
\gl +PSe1	\gl +PSe1	\gl +PSe1	\gl +PSe1	\gl +PSe1

\lf +d	\lf +ad	\lf +ed	\lf +od	\lf +öd
\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE
\alt Rag	\alt Rag	\alt Rag	\alt Rag	\alt Rag
\gl +PSe2	\gl +PSe2	\gl +PSe2	\gl +PSe2	\gl +PSe2

\lf +a	\lf +e	\lf +ja	\lf +je
\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE
\alt Rag	\alt Rag	\alt Rag	\alt Rag
\gl +PSe3	\gl +PSe3	\gl +PSe3	\gl +PSe3

\lf +nk	\lf +unk	\lf +ünk
\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE
\alt Rag	\alt Rag	\alt Rag
\gl +PSt1	\gl +PSt1	\gl +PSt1

\lf +tok	\lf +atok	\lf +otok	\lf +tek
----------	-----------	-----------	----------

\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	
\alt Rag	\alt Rag	\alt Rag	\alt Rag	
\gl +PSt2	\gl +PSt2	\gl +PSt2	\gl +PSt2	
\lf +etek	\lf +tök	\lf +ötök		
\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE		
\alt Rag	\alt Rag	\alt Rag		
\gl +PSt2	\gl +PSt2	\gl +PSt2		
\lf +uk	\lf +ük	\lf +juk	\lf +jük	\lf +k
\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE
\alt Rag	\alt Rag	\alt Rag	\alt Rag	\alt Rag
\gl +PSt3	\gl +PSt3	\gl +PSt3	\gl +PSt3	\gl +PSt3

% Possive lexicon (possessed in plural)

\lf +im	\lf +aim	\lf +eim	\lf +jaim	
\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	
\alt Rag	\alt Rag	\alt Rag	\alt Rag	
\gl +PSe1i	\gl +PSe1i	\gl +PSe1i	\gl +PSe1i	
\lf +id	\lf +aid	\lf +eid		
\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE		
\alt Rag	\alt Rag	\alt Rag		
\gl +PSe2i	\gl +PSe2i	\gl +PSe2i		
\lf +i	\lf +ai	\lf +ei		
\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE		
\alt Rag	\alt Rag	\alt Rag		
\gl +PSe3i	\gl +PSe3i	\gl +PSe3i		
\lf +ink	\lf +aink	\lf +eink	\lf +jaink	
\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	
\alt Rag	\alt Rag	\alt Rag	\alt Rag	
\gl +PSt1i	\gl +PSt1i	\gl +PSt1i	\gl +PSt1i	
\lf +itok	\lf +itek	\lf +aitok	\lf +eitek	\lf +jaitok
\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE
\alt Rag	\alt Rag	\alt Rag	\alt Rag	\alt Rag
\gl +PSt2i	\gl +PSt2i	\gl +PSt2i	\gl +PSt2i	\gl +PSt2i
\lf +ik	\lf +aik	\lf +eik	\lf +jaik	
\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	\lx POSSESSIVE	
\alt Rag	\alt Rag	\alt Rag	\alt Rag	
\gl +PSt3i	\gl +PSt3i	\gl +PSt3i	\gl +PSt3i	

% Possessive marker lexicon

\lf +é	\lf +éi
--------	---------

```

\lx POSSMARK \lx POSSMARK
\alt Rag \alt Rag
\gl +POS \gl +POSi

% Accusative lexicon

\lf +t \lf +at \lf +et \lf +ot
\lx ACCUSATIVE \lx ACCUSATIVE \lx ACCUSATIVE \lx ACCUSATIVE
\alt End \alt End \alt End \alt End
\gl +ACC \gl +ACC \gl +ACC \gl +ACC

% Illative lexicon

\lf +ba \lf +be
\lx ILLATIVE \lx ILLATIVE
\alt End \alt End
\gl +ILL \gl +ILL

% Inessive lexicon % Elative lexicon
\lf +ban \lf +ben \lf +ból \lf +ból
\lx INESSIVE \lx INESSIVE \lx ELATIVE \lx ELATIVE
\alt End \alt End \alt End \alt End
\gl +INE \gl +INE \gl +ELA \gl +ELA

% Allative lexicon

\lf +hoz \lf +hez \lf +höz
\lx ALLATIVE \lx ALLATIVE \lx ALLATIVE
\alt End \alt End \alt End
\gl +ALL \gl +ALL \gl +ALL

% Terminative lexicon % Adessive lexicon
\lf +ig \lf +nál \lf +nél
\lx TERMINATIVE \lx ADESSIVE \lx ADESSIVE
\alt End \alt End \alt End
\gl +TER \gl +ADE \gl +ADE

% Ablative lexicon % Sublative lexicon
\lf +tól \lf +től \lf +ra \lf +re
\lx ABLATIVE \lx ABLATIVE \lx SUBLATIVE \lx SUBLATIVE
\alt End \alt End \alt End \alt End
\gl +ABL \gl +ABL \gl +SUB \gl +SUB

% Superessive lexicon
\lf +n \lf +on \lf +en \lf +ön
\lx SUPERESSIVE \lx SUPERESSIVE \lx SUPERESSIVE \lx SUPERESSIVE
\alt End \alt End \alt End \alt End
\gl +SUP \gl +SUP \gl +SUP \gl +SUP

```

```

% Delative lexicon          % Dative-genitive lexicon
\lf +ról      \lf +röl      \lf +nak      \lf +nek
\lx DELATIVE  \lx DELATIVE  \lx DATIVE    \lx DATIVE
\alt End      \alt End      \alt End      \alt End
\gl +DEL      \gl +DEL      \gl +DAT      \gl +DAT

% Essive-modal lexicon     % Formal lexicon
\lf +ul       \lf +ül       \lf +ként
\lx ESSIVE    \lx ESSIVE    \lx FORMAL
\alt End      \alt End      \alt End
\gl +ESS      \gl +ESS      \gl +FOR

% Factive lexicon
\lf +vá       \lf +vé       \lf +bá       \lf +bé       \lf +cá       \lf +cé
\lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE
\alt End      \alt End      \alt End      \alt End      \alt End      \alt End
\gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC

\lf +dá       \lf +dé       \lf +fá       \lf +fé       \lf +gá       \lf +gé
\lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE
\alt End      \alt End      \alt End      \alt End      \alt End      \alt End
\gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC

\lf +há       \lf +hé       \lf +já       \lf +jé       \lf +ká       \lf +ké
\lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE
\alt End      \alt End      \alt End      \alt End      \alt End      \alt End
\gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC

\lf +lá       \lf +lé       \lf +má       \lf +mé       \lf +ná       \lf +né
\lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE
\alt End      \alt End      \alt End      \alt End      \alt End      \alt End
\gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC

\lf +pá       \lf +pé       \lf +qá       \lf +qé       \lf +rá       \lf +ré
\lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE
\alt End      \alt End      \alt End      \alt End      \alt End      \alt End
\gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC

\lf +sá       \lf +sé       \lf +tá       \lf +té       \lf +wá       \lf +wé
\lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE
\alt End      \alt End      \alt End      \alt End      \alt End      \alt End
\gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC

\lf +xá       \lf +xé       \lf +yá       \lf +yé       \lf +zá       \lf +zé
\lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE   \lx FACTIVE
\alt End      \alt End      \alt End      \alt End      \alt End      \alt End
\gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC      \gl +FAC

% Instrumental lexicon

```

\lf +val	\lf +vel	\lf +bal	\lf +bel
\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL
\alt End	\alt End	\alt End	\alt End
\gl +INS	\gl +INS	\gl +INS	\gl +INS
\lf +cal	\lf +cel	\lf +dal	\lf +del
\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL
\alt End	\alt End	\alt End	\alt End
\gl +INS	\gl +INS	\gl +INS	\gl +INS
\lf +fal	\lf +fel	\lf +gal	\lf +gel
\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL
\alt End	\alt End	\alt End	\alt End
\gl +INS	\gl +INS	\gl +INS	\gl +INS
\lf +hal	\lf +hel	\lf +jal	\lf +jel
\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL
\alt End	\alt End	\alt End	\alt End
\gl +INS	\gl +INS	\gl +INS	\gl +INS
\lf +kal	\lf +kel	\lf +lal	\lf +lel
\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL
\alt End	\alt End	\alt End	\alt End
\gl +INS	\gl +INS	\gl +INS	\gl +INS
\lf +mal	\lf +mel	\lf +nal	\lf +nel
\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL
\alt End	\alt End	\alt End	\alt End
\gl +INS	\gl +INS	\gl +INS	\gl +INS
\lf +pal	\lf +pel	\lf +qal	\lf +qel
\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL
\alt End	\alt End	\alt End	\alt End
\gl +INS	\gl +INS	\gl +INS	\gl +INS
\lf +ral	\lf +rel	\lf +sal	\lf +sel
\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL
\alt End	\alt End	\alt End	\alt End
\gl +INS	\gl +INS	\gl +INS	\gl +INS
\lf +tal	\lf +tel	\lf +wal	\lf +wel
\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL
\alt End	\alt End	\alt End	\alt End
\gl +INS	\gl +INS	\gl +INS	\gl +INS
\lf +xal	\lf +xel	\lf +yal	\lf +yel
\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL	\lx INSTRUMENTAL
\alt End	\alt End	\alt End	\alt End
\gl +INS	\gl +INS	\gl +INS	\gl +INS

```

\lf +zal          \lf +zel
\lx INSTRUMENTAL \lx INSTRUMENTAL
\alt End          \alt End
\gl +INS          \gl +INS

% Sociative lexicon          % Causal-final lexicon
\lf +stul          \lf +stül          \lf +ért
\lx SOCIATIVE      \lx SOCIATIVE      \lx CAUSAL
\alt End          \alt End          \alt End
\gl +SOC          \gl +SOC          \gl +CAU

% Distributive lexicon
\lf +nként          \lf +onként          \lf +enként          \lf +önként
\lx DISTRIBUTIVE    \lx DISTRIBUTIVE    \lx DISTRIBUTIVE    \lx DISTRIBUTIVE
\alt End          \alt End          \alt End          \alt End
\gl +DIS          \gl +DIS          \gl +DIS          \gl +DIS

% Temporal lexicon          % Distributive temporal lexicon
\lf +kor          \lf +onta          \lf +anta          \lf +ente
\lx TEMPORAL      \lx DISTEMP      \lx DISTEMP      \lx DISTEMP
\alt End          \alt End          \alt End          \alt End
\gl +TEM          \gl +DIT          \gl +DIT          \gl +DIT

```

B Tagguppsättning

För att markera suffixen har ett antal taggar använts. Nedan finns samtliga taggar listade (i bokstavsordning för att underlätta sökandet). Tillsammans med taggarna finns suffixnamnet, exempel och översättning.

Tagg	Namn	Exempel	Översättning
+ABL	ablativ	-tól	'från, av'
+ACC	ackusativ	-t	<i>objektsform</i>
+ADE	adessiv	-nál	'hos, vid'
+ALL	allativ	-hoz	'till'
+CAU	kausalfinal	-ért	'för...skull'
+DAT	dativ	-nak	'åt'
+DEL	delativ	-ról	'ned från, om'
+DIS	distributiv	-nként	'per, på, om'
+DIT	distributiv-temporalis	-nta	'per, om'
+ELA	elativ	-ból	'ut ur'
+ESS	essiv-modal	-ul	'såsom'
+FAC	translativ-faktiv	-vá	'förändring till'
+FOR	essiv-formalis	-ként	'såsom'
+ILL	illativ	-ba	'in i'
+INE	inessiv	-ban	'i'
+INS	instrumental-komitativ	-val	'med, tillsammans med'
+NOM	nominativ	-	<i>grundform</i>
+PL	pluralis	-k	<i>pluralform</i>
+POS	possessivmärke, det ägda i singular	-é	<i>genitiv</i>
+POSi	possessivmärke, det ägda i plural	-éi	<i>genitiv</i>
+PSe1	possessiv, första person singular, det ägda i singular	-m	'min'
+PSe1i	possessiv, första person singular, det ägda i plural	-im	'mina'
+PSe2	possessiv, andra person singular, det ägda i singular	-d	'din'
+PSe2i	possessiv, andra person singular, det ägda i plural	-id	'dina'
+PSe3	possessiv, tredje person singular, det ägda i singular	-a	'hans/hennes'
+PSe3i	possessiv, tredje person singular, det ägda i plural	-i	'hans/hennes'
+PSt1	possessiv, första person plural, det ägda i singular	-nk	'vår'
+PSt1i	possessiv, första person plural, det ägda i plural	-ink	'våra'
+PSt2	possessiv, andra person plural, det ägda i singular	-tok	'er'
+PSt2i	possessiv, andra person plural, det ägda i plural	-itok	'era'
+PSt3	possessiv, tredje person plural, det ägda i singular	-uk	'deras'
+PSt3i	possessiv, tredje person plural, det ägda i plural	-ik	'deras'
+SOC	instrumental-sociativ	-stul	'tillsammans med'
+SUB	sublativ	-ra	'upp på, till'
+SUP	superessiv	-n	'på'
+TEM	temporal	-kor	'vid'
+TER	terminativ	-ig	'fram till'