De invloed van morfologie op inductieve grafeem-foneemconversie

Dieter Van Uytvanck
studentnummer 0244325
10 februari 2006
Begeleiders: Erwin Marsi en Joop Kerkhoff

Voorwoord

Graag zou ik iedereen willen bedanken die heeft bijgedragen tot de verwezenlijking van deze bachelorscriptie , in het bijzonder gaat mijn dank uit naar:

- Erwin Marsi, voor de goede begeleiding, de tips, het aanleveren en bespreken van ideeën.
- Antal van den Bosch, voor de hulp en expertise in verband met MBMA.
- Ko van der Sloot, voor de aangebrachte verbeteringen aan TiMBL.
- De mensen uit Nijmegen die voor de nodige ondersteuning en begeleiding zorgden: Joop Kerkhoff, Peter-Arno Coppen en Lou Boves.
- Mijn familie, voor haar steun en begrip.

Samenvatting

In dit onderzoek wordt nagegaan of het gebruik van morfologische features (naast grafemische features) een invloed heeft op de prestaties van inductieve grafeem-foneemconversie. Daartoe is binnen NeXTeNS de TreeTalk-module ook getraind met morfologische samenstellingsfeatures. De morfologische analyse werd uitgevoerd met behulp van MBMA, een geheugengebaseerde parser. Uit 10-fold cross-validation blijkt dat hoewel de morfologie ook gebruikt wordt bij de grafeem-foneemomzetting er geen significant verschil bestaat tussen de TreeTalk-variant met en zonder de informatie over samenstellingen.

Inhoudsopgave

1 Inleiding

Spraaksynthese is al geruime tijd de kinderschoenen ontgroeid. Ook voor het Nederlands zijn verschillende functionerende tekst-naar-spraaksystemen (TTS[1]) ontwikkeld. Eén daarvan is NeXTeNS [Marsi and Kerkhoff, 2003], een project dat verschillende al bestaande componenten en lexica (KunTTS, FonPars , TreeTalk, etc. [Kerkhoff et al., 1998, Kerkhoff and Rietveld, 1994, Daelemans and van den Bosch, 1993]) integreert tot een spraaksynthese-systeem dat vrij te gebruiken is voor niet-commerciële doeleinden. Het bestaat – zoals alle TTS-toepassingen – uit verschillende modules [Marsi and Kerkhoff, 2002]:

1. De tekst opdelen in tokens, herkenning van woorden en leestekens.
2. Woordsoortherkenning (POS-tagging)
3. Syntactische analyse
4. Omzetting van tokens naar woorden, waarbij speciale patronen (getallen en dergelijke) naar hun uitspraak-equivalent omgezet worden.
5. Prosodie-generatie
6. Grafeem-foneemconversie
7. Duurbepaling: voor elk foneem bepalen hoe lang het duurt.
8. Berekening van F0 in functie van de tijd
9. Golfvorm-synthese gebaseerd op difonen.

In het verdere verloop van dit onderzoek zullen we de aandacht richten op stap 6, de grafeem-foneemconversie.

1.1 Methodes voor grafeem-foneemomzetting

De meest eenvoudige manier om de letters van een woord in fonemen om te zetten is natuurlijk door de uitspraak van woorden op te zoeken in een lexicon. Dat is ook de standaardmethode die NeXTeNS toepast. Als een woord niet voorkomt in het uitspraakwoordenboek moet er echter overgegaan worden op een alternatieve methode. Het zijn die technieken die het onderwerp uitmaken van dit onderzoek en hieronder kort besproken worden.

1.1.1 Connectionistische modellen

In [Sejnowski and Rosenberg, 1987] wordt een methode voorgesteld om een neuraal netwerk te trainen op grafeem-foneemomzetting voor het Engels. Dergelijke backpropagationmodellen hebben over het algemeen het voordeel dat ze goed zijn in het veralgemenen van impliciete regelmatigheden, terwijl toch ook de uitzonderingen behouden kunnen blijven.

Het systeem werd onderworpen aan een test voor het Nederlands in [Daelemans and van den Bosch, 1993]. In vergelijking met andere (tabelgebaseerde en inductieve) methodes levert het mindere prestaties.

[1] Text-to-Speech

1.1.2 Tabelgebaseerde modellen

Een zeer basale conversiemanier bestaat erin om eerst een tabel te vullen met trainingsdata. Per geval wordt een letter, zijn contextletters en de uitspraak van de letter in kwestie opgeslagen. Eventueel kan de tabel ook gecomprimeerd worden. Dat houdt in dat voor elke letter alleen de minimale context bewaard wordt die nodig is om ondubbelzinnig de overeenkomstige uitspraak te vinden. Zo zal het grafeem "x" in het Nederlands zelfs helemaal geen context vereisen: in alle gevallen wordt het uitgesproken als /ks/. Voor een letter als "e" daarentegen, die op vele manieren uitgesproken kan worden, is een ruime context dan weer onmisbaar.

Uit evaluatie van deze methode in onder andere [Daelemans and van den Bosch, 1993] blijkt duidelijk dat hiermee zeer goede resultaten behaald worden[2].

1.1.3 Regelgebaseerde modellen

De grafeem-foneemomzetting van NeXTeNS kan op twee manieren uitgevoerd worden. Eén daarvan is FonPars [Kerkhoff and Rietveld, 1994], een regelgebaseerd systeem. Dit maakt expliciet gebruik van fonologische regels, zoals beschreven in [Chomsky and Halle, 1968]. In eerste instantie worden alle grafemen in een standaardfoneem omgezet ("b" wordt /b/ bijvoorbeeld voor het Nederlands). Vervolgens worden contextafhankelijke herschrijfregels op fonemen toegepast. Zo zal de "b" aan het einde van "web" als een /p/ (i.e. stemloos) uitgesproken worden.

Voor een kwalitatieve evaluatie van regelgebaseerde modellen verwijzen we naar sectie 1.1.5.

1.1.4 Probabilistische modellen

De eerder genoemde regelgebaseerde modellen kunnen ook per herschrijfregel uitgebreid worden met informatie over de waarschijnlijkheid dat een regel toegepast moet worden. Dergelijke probabilistische grafeem-foneemconversie is met succes toegepast in onder andere [Tarsaku et al., 2001] voor het Thaïs met een accuraatheid van 72.78%, een resultaat dat de scores van regelgebaseerde en inductieve modellen voor die taal overstijgt. Niettemin is dit resultaat waarschijnlijk toe te schrijven aan de specifieke vereisten die het Thaïs op dit gebied stelt.

1.1.5 Inductieve modellen

Als laatste zijn er ook de modellen die steunen op Instance-Based Learning (IBL). Zoals de naam al aangeeft gaat het hier om zelflerende systemen die – naar analogie met de exemplar-theorie uit de psychologie – zich baseren op gelijkenissen met bekende prototypes. Voor de gelijkenissen wordt i.c. naar de uit te spreken letter (en zijn context) gekeken. De gekozen uitspraak is uiteindelijk die van het prototype die de ingevoerde instantie het meeste benadert.

Een voorbeeld van een IBL-systeem is TiMBL [Daelemans et al., 2004]. Als TiMBL gebruikt wordt in de context van grafeem-foneemconversie wordt krijgt het de naam TreeTalk [Busser, 1998]. TreeTalk bepaalt dus welk foneem er bij een letter hoort. In haar scriptie nam Nanneke Konings [Konings, 2003] naast TreeTalk ook FonPars onder de loep en ze vergeleek ze deze aan de hand van 4 criteria:

[2] Een accuraatheid van 95.1%.

- klankomzetting
- klemtoontoekenning
- syllabificatie
- samenstellingsgrenzen

Ze komt tot de conclusie dat TreeTalk op alle vlakken beter presteert dan FonPars. De voordelen van geheugengebaseerde grafeem-foneemomzetting komen bijvoorbeeld expliciet naar boven bij leenwoorden die de "klassieke" uitspraakregels niet volgen. Anderzijds gaat deze methode af en toe ook in de fout, zo worden er soms 2 primaire klemtonen binnen één woord gelegd. Alles wel beschouwd is het verschil tussen beide methodes van die aard dat het logisch is om vervolgonderzoek toe spitsen op grafeem-foneemomzetting met TreeTalk.

1.2 TreeTalk

NeXTeNS beschikt tijdens het tekst-naar-spraakproces standaard niet over de morfologische opbouw van de invoerwoorden. Nochtans kan morfologie een factor zijn die op 3 manieren kan bijdragen tot een verbeterde grafeem-foneemomzetting:

- Een verbeterde syllabificatie: TreeTalk splitst het woord autofabriek als au–tofa–briek [Konings, 2003]. Met kennis over de achterliggende samenstelling (auto–fabriek) zou de correcte splitsing au–to–fa–briek gevonden kunnen worden.
- Als een samenstelling als dusdanig herkend wordt, kan er ook een correcte primaire en secundaire klemtoon geplaatst worden. Zo wordt "voetbalvereniging" niet als samenstelling door NeXTeNS herkend, waardoor de secundaire klemtoon (voetbalveréniging) ontbreekt.
- De mogelijkheid bestaat dat een zelflerend systeem als TreeTalk tijdens de leerfase voordeel kan halen uit –naast de zuivere grafeeminvoer– contextinformatie over de morfologie van het om te zetten woord. Zo wordt klinkt de /n @ t j @/[3] van "mannetje" anders dan de /n E t j @/ van "muggennetje", hoewel beiden op dezelfde manier geschreven zijn. Het systeem zou in dit geval met succes de ambiguïteit kunnen wegwerken omdat het weet dat het de eerste "e" van "netje" als deel van een samenstelling op een bepaalde manier moet uitspreken.

Er zijn duidelijk mogelijkheden te over om de grafeem-foneemomzetting te proberen verbeteren met behulp van morfologische analyse. Omwille van praktische redenen en de beperkte tijd die voorhanden was voor dit onderzoek is de keuze gemaakt om alleen de invloed na te gaan van morfologie-informatie op de fonemisatie.

[3] Alle fonetische transcripties zijn genoteerd met behulp van de CGN-foneemset[Gillis, 2001].

2 Probleem- en doelstelling

2.1 Werking van TreeTalk

Opmerkelijk genoeg bevat TreeTalk geen linguïstische kennis in de vorm van fonologische regels. Aan de hand van trainingsmateriaal "leert" het als het ware met welk foneem een grafeem overeenkomt. Deze inductieve manier van werken heeft al op verschillende vlakken in de taal- en spraaktechnologie zijn kwaliteiten bewezen. Zo is het onder meer met succes gebruikt voor klemtoontoekenning , syllabificatie en woordsoortherkenning [van den Bosch and Daelemans, 2005].

Hoe werkt grafeem-foneemomzetting met TreeTalk nu precies?

2.1.1 Leerfase

Om tot een inductieve classificatie te komen moet het systeem natuurlijk eerst getraind worden. Dat gebeurt door een grote hoeveelheid voorbeelden ("instances") in te voeren. Een instance bestaat uit een te classificeren element – in dit geval een grafeem – samen met de omliggende elementen en de juiste klasse (i.c. een foneem). De opslag van de trainingsdata kan op verschillende manieren gebeuren. Alle featurevectoren kunnen volledig bewaard worden of er kan verdere abstractie en pruning plaatsvinden, bijvoorbeeld door de constructie van beslissingsbomen. Details hierover zijn te vinden in [Daelemans et al., 2004]; voor de training van TreeTalk hebben we ervoor gekozen om de instances op te slaan als IG[4]-bomen. Deze methode is sneller en gebruikt minder geheugen dan IB1[5], de standaard voor TiMBL. Als keerzijde levert het waarschijnlijk suboptimale resultaten. Er is dus sprake van een klassieke trade-off. In een applicatie als NeXTeNS is snelheid echter van groot belang, vandaar de keuze voor IG.
Het spreekt daarnaast voor zich dat er veel trainingsgegevens vereist zijn om tot een betrouwbaar resultaat te komen.

2.1.2 Classificatie

In eerste instantie wordt het te classificeren token omgezet in een featurevector. Die vormt een soort "venster" (sliding window) dat buiten de om te zetten letter (het centrum van het venster) ook de omringende letters bevat (de context). Zo zal de eerste "e" in "alpenwei" voorgesteld worden door de featurevector [_ a l p **e** n w e i], waarbij "_ a l p"[6] en "nwei" als context beschouwd worden.

De volgende stap in het classificatieproces bestaat erin om de featurevector te vergelijken met het aangeleerde materiaal (de "instance base"). Hierbij zijn verschillende scenario's mogelijk:

De featurevector bestaat in de instance base In dit geval wordt gewoon de corresponderende klasse uit de instance base gekozen. In ons voorbeeld zou de klasse van de e sjwa (@) zijn. Komen er voor een bepaalde featurevector meerdere klasses voor,

[4]Information Gain
[5]Instance Base 1
[6]De underscore _ is een dummy-symbool voor lege elementen van de featurevector.

Tabel 1: grafeem-foneemomzetting voor "alpenwei"

instantienummer					focus-feature					classificatie
1	_	_	_	_	**a**	l	p	e	n	**A**
2	_	_	_	a	**l**	p	e	n	w	**l**
3	_	_	a	l	**p**	e	n	w	e	**p**
4	_	a	l	p	**e**	n	w	e	i	**@**
5	a	l	p	e	**n**	w	e	i	_	**-**
6	l	p	e	n	**w**	e	i	_	_	**w**
7	p	e	n	w	**e**	i	_	_	_	**-**
8	e	n	w	e	**i**	_	_	_	_	**E+**

dan kiest het systeem die klasse die tijdens de training het meest voorkwam bij deze featurevector.

De featurevector bestaat niet in de instance base De klasse van de meest gelijkende featurevector is in dat geval het gepaste antwoord. Hoe de gelijkenis (of anders gesteld: de afstand) tussen 2 featurevectoren bepaald wordt, valt onder meer te lezen in [Daelemans et al., 2004].

Uiteindelijk zal de TreeTalk-classificatie van het woord "alpenwei" er uitzien als in tabel 1. De laatste kolom geeft aan op welk foneem een grafeem afgebeeld wordt.

De vergelijking van featurevectoren kan ook aangepast worden aan het relatieve belang van een feature op de classificatie. In zo'n geval spreek men van featureweging. Bij onze experimenten is voor elk feature de gemiddelde invloed op de classificatie berekend (de "information gain ratio", zie o.a. [Daelemans and van den Bosch, 1993]). Deze gegevens zijn vervolgens gebruikt om de features te wegen.

2.1.3 Prestaties

Volgens [Konings, 2003] zet TreeTalk 76.0% van alle aangeboden woorden[7] volledig correct om in fonemen. Dat is opmerkelijk beter dan FonPars, de regelgebaseerde grafeem-foneemomzetter, die 54.4% van alle woorden foutloos omzet.

Het meeste gaat TreeTalk in de fout bij de fonemen /x/ en /G/. Zo wordt "geboortes" in /x@bort@s/ omgezet in plaats van het correcte /G@bort@s/. Aangezien beide vormen vrij sterk op elkaar gelijken geeft de onderlinge verwarring geen aanleiding tot goed hoorbare uitspraakfouten. Hetzelfde geldt voor het al dan niet fonemiseren van de tussen-n in samenstellingen, zoals bij "varkenshouder". Een deel van deze fouten is te wijten aan het voorkomen van foute fonetische transcripties in Kunlex, het trainingslexicon voor TreeTalk.[8]

Ernstiger is het als het grafeem "e" verkeerd uitgesproken wordt. Dit blijkt – nog steeds volgens [Konings, 2003] – de enige vaak voorkomende hoorbare fout van TreeTalk te zijn. Het geeft aanleiding tot het uitspreken van "zonnepetje" als /zOn@p@tj@/. Als alleen de ernstige (i.e. goed waarneembare) fouten in rekening gebracht worden genereert TreeTalk voor 93.9% van de woorden een foutloze fonemische transcriptie.

[7] Er is in het genoemde experiment geen gebruik gemaakt van 10-fold cross validation maar van een aparte lijst met samengestelde woorden.

[8] In sommige gevallen is bijvoorbeeld geen rekening gehouden met assimilatie, waardoor "fietsverkeer" ten onrechte als /fitsv@rker/ en niet als /fitsf@rker/ getranscribeerd is.

Daarnaast valt op te merken dat TreeTalk goed scoort bij leenwoorden. Dat valt uiteraard te verklaren door de naïviteit die gepaard gaat met inductie: als er maar voldoende leenwoorden in het trainingsmateriaal zitten, zal het systeem net zo goed de anderstalige uitspraak leren.

2.1.4 Outputfilter

Andere factoren die de grafeem-foneemomzetting bepalen zijn klemtoon en syllabificatie. Hiervoor produceert TreeTalk regelmatig onmogelijke uitspraakpatronen, zoals twee primaire klemtonen binnen een woord of lettergrepen die niet voorkomen in het Nederlands. Aangezien deze problemen vrij eenvoudig opgelost kunnen worden met heuristische regels, is er bij het ontwerp van NeXTeNS een soort post-processor geimplementeerd die deze foutieve output kan aanpassen aan de hand van een beperkt aantal correctieregels. Op zich vallen syllabificatie en klemtoontoekenning niet binnen het bestek van dit onderzoek, maar omdat beiden invloed uitoefenen op de grafeem-foneemomzetting (een beklemtoonde "e" kan bijvoorbeeld nooit als /@/ uitgesproken worden) zal bij de verdere experimenten telkens ook een resultaat mét outputfilter vermeld worden.

2.2 Werking van MBMA

Een computermatige morfologische analyse maken is geen eenvoudige taak. Uit het onderzoek van [Barton et al., 1987] blijkt zelfs dat morfologische analyse op 2 niveaus (oppervlakte- en dieptestructuur) een NP-compleet probleem is. Traditioneel is veel onderzoek gedaan naar het gebruik van finite state transducers. Een veel voorkomende hinderpaal daarbij is dat het manueel opstellen van regels voor morfologische analyse tijdrovend en moeilijk is.

In [van den Bosch and Daelemans, 1999] wordt daarom onder de naam MBMA als alternatief voorgesteld om deze taak te bekijken als een geheugengebaseerde classificatieopdracht. Volledig naar analogie met de eerder besproken TreeTalk (en eveneens gebaseerd op TiMBL) wordt voor elke letter van een woord nagegaan of het om een morfologische grens gaat en indien dat het geval is van wat voor een morfeem er sprake is. De auteurs rapporteren een precisie en recall van 84% voor woordenboekwoorden en naar schatting 93% voor vrije tekst.

Ter illustratie is in tabel 2 een dergelijke analyse opgenomen voor het woord $[[Alpen]_N[wei]_N]_N$. De codes voor de morflogische klasses die toegekend worden zijn overgenomen uit CELEX [Baayen et al., 1993]. Zo betekent de klasse van de "a" in alpenwei, N+RA>a, dat dit morfeem van oorsprong een nomen is (N) en dat de schrijfwijze aangepast is (de "A" uit Alpen wordt "a" in alpenwei). Voor de opzet van het verdere onderzoek is vooral de informatie over samenstellingen die MBMA genereert van belang.

Tabel 2: morfologische analyse voor "alpenwei"

instantienummer					focus-feature					classificatie
1	_	_	_	_	a	l	p	e	n	N+RA>a
2	_	_	_	a	**l**	p	e	n	w	0
3	_	_	a	l	**p**	e	n	w	e	0
4	_	a	l	p	**e**	n	w	e	i	0
5	a	l	p	e	**n**	w	e	i	_	0
6	l	p	e	n	**w**	e	i	_	_	N
7	p	e	n	w	**e**	i	_	_	_	0
8	e	n	w	e	**i**	_	_	_	_	0/e

2.3 Onderzoeksopzet

2.3.1 Doelstelling

Uit [Booij and van Santen, 1998] blijkt dat onder de 89.000 gelede woorden uit CELEX[9] er ruim 63.000 samenstellingen zijn. Het merendeel (55.000) daarvan bestaat dan nog eens uit nominale samenstellingen. De samenstelling – en in het bijzonder de nominale – is dus voor het Nederlands een uiterst productief proces. Het is daardoor onmogelijk om al deze samenstellingen op te nemen in een uitspraaklexicon voor Nederlandse tekst-naar-spraaksystemen. Vandaar het enorme belang van een grafeem-foneemomzetter die buiten een basislexicon om ook overweg kan met (onder andere nieuwe) samenstellingen.

2.3.2 Probleemstelling

Omdat de inductie achter TreeTalk en MBMA niet op een expliciet regelsysteem gebaseerd is bleek het aantrekkelijk en relatief eenvoudig realiseerbaar om beide systemen te combineren. In plaats van te formaliseren welke invloed morfologie op fonologie heeft volstond het om TiMBL te trainen met gecombineerde instanties. Die bevatten naast de grafeem-foneemafbeelding ook morfologische informatie. De idee hierachter is dat de uitgebreide TreeTalk zelf generalisaties zal kunnen afleiden uit het trainingscorpus en dus indirect de invloed zal aanleren van morfologie op fonologie en dankzij die bredere "kennnis" tot betere resultaten zal komen voor de grafeem-foneemconversie.

Formeel gezien onderzoeken we volgende nulhypotheses:

Hypothese 1 Inductieve grafeem-foneemomzetting presteert even goed mét als zónder extra morfologische informatie tijdens de training en classificatie.

Hypothese 2 De grafeem-foneemconversie met morfologie maakt dezelfde fouten als die zonder morfologie.

[9]De Nederlandse versie van CELEX bevat in het totaal 381.292 woordvormen, wat overeenkomt met 124.136 lemmata.

3 Experimenten en resultaten

3.1 Materiaal

3.1.1 Trainingscorpus TreeTalk

Voor de training van TreeTalk is net zoals bij NeXTeNS gebruik gemaakt van het Kunlex-lexicon. Dat bevat 316.996 woorden. Woorden die speciale karakters bevatten zoals trema's en koppeltekens (bijvoorbeeld A'dam en A-bom) werden gemakshalve niet gebruikt bij het experiment. Na deze filtering telde het lexicon nog 311.836 woorden.

Overigens bevatte deze lijst nog steeds dezelfde fouten als beschreven in [Konings, 2003]. Het gaat dan over:

- meerdere primaire klemtonen in een woord, bijvoorbeeld
  ("beginjaren" nil ((((b @) 0) ((x I n) 1)) (((j a) 1) ((r @ n) 0))))
- een verkeerde toepassing van assimilatieregels[10], bijvoorbeeld
  ("rupsband" nil ((((r Y b z) 1)) (((b A n t) 2))))
- willekeurige fouten, bijvoorbeeld
  ("taxibedrijf" nil ((((t A k) 1)) (((b @) 0) ((d r E+ f) 2))))

Een mogelijke oplossing hiervoor zou erin kunnen bestaan om Kunlex te vergelijken met Celex [Baayen et al., 1993] en zo de fouten op te sporen, of eenvoudigweg om Celex te gebruiken als lexicon. Bij gebrek aan tijd is er voor gekozen om de volgende experimenten toch op Kunlex [Kerkhoff et al., 1998] te baseren.

3.1.2 Trainingscorpus MBMA

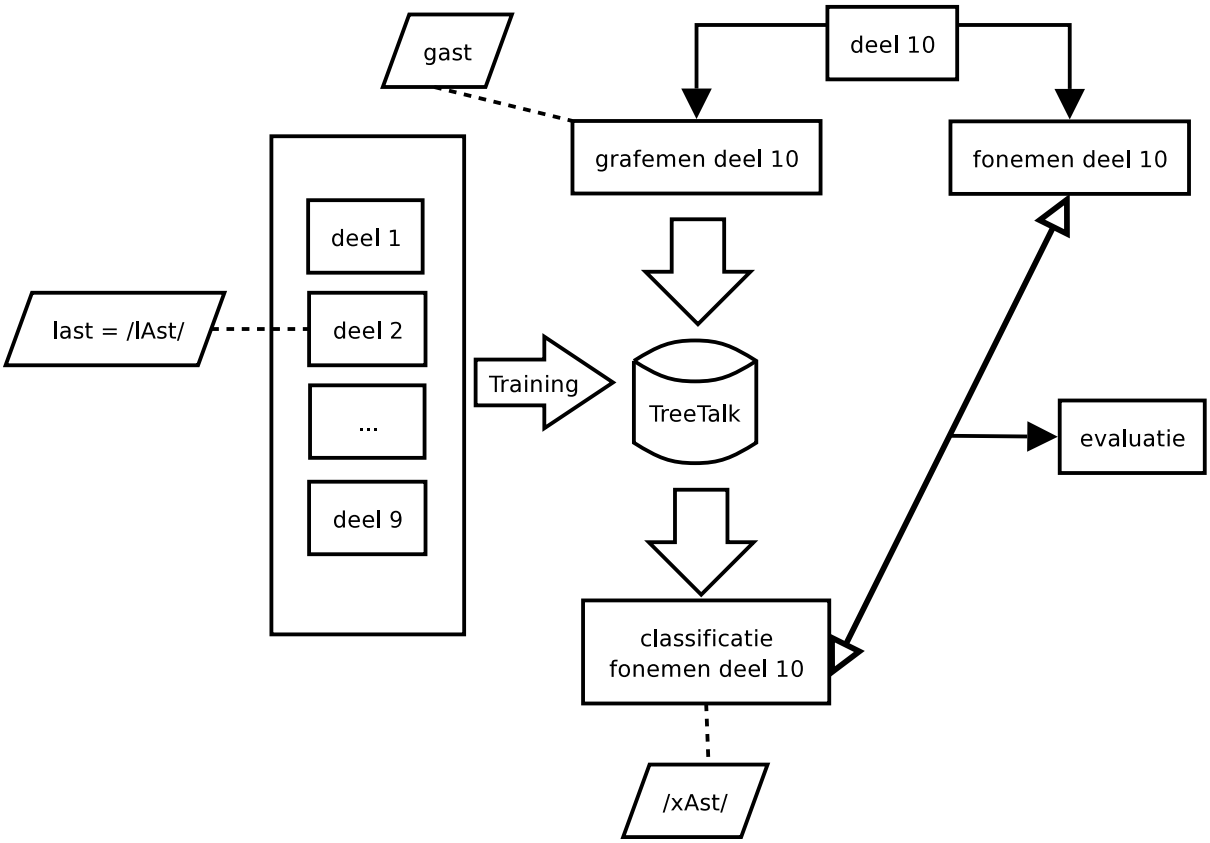
Naar analogie met [van den Bosch and Daelemans, 1999] hebben we als trainingsmateriaal voor MBMA gebruik gemaakt van Celex, wat neerkomt op 247.415 volledig morfologisch geanalyseerde woorden. De training gebeurde met een linker- en rechtercontext van 6 karakters.

3.2 Procedure

3.2.1 De originele TreeTalk: het referentiepunt

Om een goed ijkpunt vast te stellen voor de vergelijking van Treetalk met en zonder beschikking over morfologie hebben we in eerste instantie de klassieke versie van TreeTalk getraind en geëvalueerd. Dit gebeurde volgens de "10-fold cross-validation"-methode. Het lexicon werd in 10 gelijke delen gesplitst. Telkens werd TreeTalk getraind met 9 delen daarvan. Daarna volgde de evaluatie van de resterende 10%. Door deze procedure 10 maal te herhalen – zodat elke groep precies 1 keer geëvalueerd werd – was het mogelijk om een betrouwbaar beeld te scheppen van de prestaties van dit zelflerende systeem, zoals dit eerder beschreven werd in [Weiss and Kulikowski, 1991]. Figuur 1 geeft dit proces schematisch weer. Zowel de training als evaluatie

[10] Deze fout wordt in het TTS-systeem al verbeterd door toepassing van postlexicale regels.



Figuur 1: Experimentele opzet voor de evaluatie van TreeTalk. Stippellijnen geven voorbeelden aan.

gebeurden met een linker- en rechtercontext van 4 grafemen, naar analogie met de bestaande NeXTeNS.

Een bijkomend voordeel van deze procedure is dat er zeker geen verschillen bestaan wat notatie- en transcriptieconventies betreft tussen het trainings- en het testmateriaal.

3.2.2 TreeTalk met MBMA: de uitbreiding

In deze fase voegden we naast het bestaande trainingsmateriaal van TreeTalk ook een morfologische featurevector toe aan elke instantie van de trainingsset. Omdat we vooral geïnteresseerd waren in een potentiële verbetering bij nominale samenstellingen hebben we er voor gekozen om met een soort vereenvoudigde featurestructuur te werken om de samenstellingsgrenzen aan te geven.

Om daartoe te komen wordt er gestart met een morfologische analyse met behulp van MBMA. Stel dat we het woord "dakpan" $[[dak]_N[pan]_N]_N$willen classificeren (of als trainingsmateriaal willen gebruiken). MBMA zal dan achtereenvolgens de volgende klasses opleveren: N, 0, 0, N, 0, 0. Een correct resultaat, want zowel "dak" als "pan" vormen de nominale leden van deze samenstelling.

Dit resultaat wordt vervolgens vereenvoudigd tot de bitvector [0,0,0,1,0,0]. De "p" wordt een 1 omdat deze het begin is van een nieuw lid van een samenstelling (i.e. "pan"). De eerste letter van een woord wordt nooit als samenstellingsonderdeel gemarkeerd. Het ontbreken van een linkercontext in de grafeemvector geeft namelijk al

aan dat het om een woordbegin gaat. Als dit een effect heeft op de fonemisatie, hoeft het geen twee keer aangegeven te worden.

De samenstellingen[11] uit tabel 3 worden gemarkeerd in de bitvector. Dit is een beperkte selectie die desalniettemin de meest frequente samenstellingen omvat. Als het prefix voor een deel van de samenstelling "ver-", "be-" of "ge-" is, dan gaat het naar grote waarschijnlijkheid om een afleiding en wordt de samenstellingsbit dus op 0 gezet.

Tabel 3: Samenstellingsgrenzen die gemarkeerd werden voor de verrijking van de TreeTalk instanties.

Celex symbool	betekenis	voorbeeld
A	adjectief	probaat
N	nomen	pan
N+D	nomen, deletie aan het einde van het voorgaande lid	aardappelen (= aarde + appelen, deletie van de e)
V	verbum	werken
V_V*	een samenstelling van werkwoorden die tot een nieuw werkwoord leidt	trekkebekken

In een volgende fase wordt de informatie uit deze bitvector gecombineerd met de grafemen die een woord vormen tot één featurevector. De context van de grafemen (4 elementen links en rechts) blijft ongewijzigd, voor de samenstellingen hebben we gekozen voor een venster van 5 elementen (dus een linker- en rechtercontext van 2 elementen). De laatste contextbreedte is arbitrair gekozen maar zal zeker de belangrijkste (i.e. centrale) features bevatten, zo leert de ervaring met TiMBL.

In het totaal ontstaat dus een featurevector van 15 elementen: 9 voor de grafemen, 5 voor de samenstellings-bitvector en 1 voor het foneem (de toegekende klasse). Een voorbeeld van dergelijke vectoren voor het woord "dakpan" is te vinden in tabel 4.

Voor de evaluatie van dit classificatiesysteem hebben we eveneens 10-fold cross validation toegepast[12]. Een schematische weergave hiervan is te vinden in figuur 2.

Om leesbaarheidsoverwegingen zullen we in de verdere tekst verwijzen naar de combinatie van TreeTalk en MBMA met de term MTreeTalk (tegenover de "klassieke" TreeTalk).

3.2.3 Evaluatie

Om de prestaties van TreeTalk in de eerder besproken verschillende uitvoeringen te meten hebben we verschillende maten gebruikt:

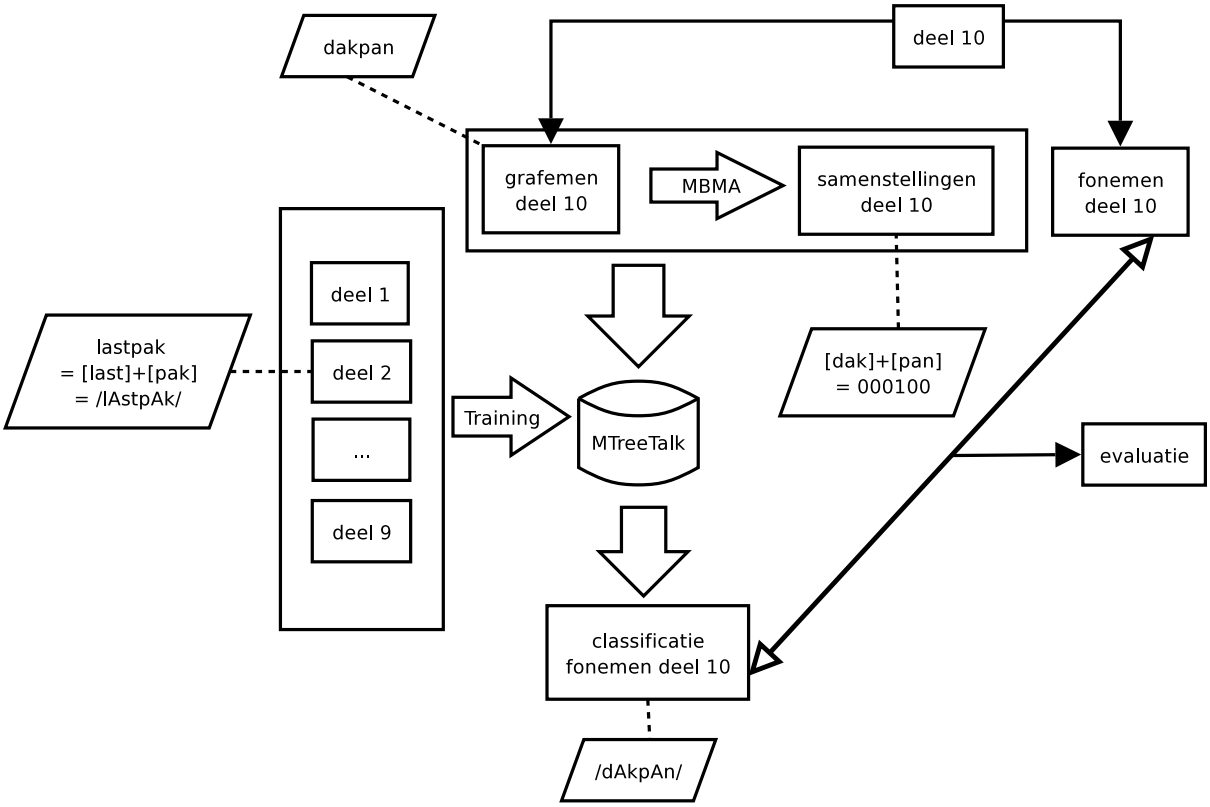
- De Word Error Rate (WER), de verhouding van het aantal woorden met minimaal 1 fout in de fonemische transcriptie tot het totaal aantal woorden.

[11] De symbolen uit Celex komen overeen met de classificatie door MBMA.

[12] Volledigheidshalve: de training van MBMA zelf is om praktische overwegingen niet volgens 10-fold cross validation verlopen. Hoewel dit zo goed als zeker geen invloed heeft op de resultaten zou het de methodologische zuiverheid wel ten goede komen.

Tabel 4: Voorbeelden van grafeem-featurevectoren in combinatie met bitvectoren die de samenstellingen aangeven.

featurevector														classificatie
grafemen									samenstelling					fonemen
_	_	_	_	**d**	a	k	p	a	_	_	**0**	0	0	**d**
_	_	_	d	**a**	k	p	a	n	_	0	**0**	0	1	**A**
_	_	d	a	**k**	p	a	n	_	0	0	**0**	1	0	**k**
_	d	a	k	**p**	a	n	_	_	0	0	**1**	0	0	**p**
d	a	k	p	**a**	n	_	_	_	0	1	**0**	0	0	**A**
a	k	p	a	**n**	_	_	_	_	1	0	**0**	0	0	**n**



Figuur 2: Experimentele opzet voor de evaluatie van MTreeTalk. Stippellijnen geven voorbeelden aan.

- De Phoneme Error Rate (PER), de verhouding van het aantal foutieve fonemen ten opzichte van het totaal aantal fonemen.
- De bedoeling was ook een overzicht te krijgen van de aard van de voorkomende fouten en de eventuele overlap daarin tussen de 2 methodes. Daarom hebben we voor de WER ook berekend hoeveel procent van de fouten bij beide methodes voorkomen. Met andere woorden: voor welke woorden gaan beide methodes de mist in?

De WER is zeer eenvoudig te bepalen, twee fonemische transcripties komen immers overeen of niet. Voor de bepaling van de PER hebben we gekozen om de minimal string distance [Wagner and Fischer, 1974] te bepalen tussen de twee versies. Daarbij kregen inserties, deleties en substituties allen hetzelfde gewicht. Voor de uiteindelijke uitspraak kunnen alledrie deze fouten immers leiden tot een zelfde mate van verwarring.

Door het gebruik van de 10-fold cross validation-procedure hebben we de PER en WER bepaald voor elk van de tien testsets. Die gegevens zijn daarna gebruikt voor het uitvoeren van een gepaarde t-toets om na te gaan of de gemiddeldes voor TreeTalk en MTreeTalk significant verschillen. Met andere woorden: er wordt nagegaan of het beschouwen van morfologie een effect heeft op de grafeem-foneemconversie.

Er bestaat binnen NeXTeNS, zoals eerder vermeld, een outputfilter dat onmogelijke foneemsequenties probeert te corrigeren. Volledigheidshalve hebben we bij het uitvoeren van de experimenten steeds ook een keer getest mét dit filter.

3.3 Resultaten

3.3.1 Word Error Rate

De Word Error Rate voor alle tien test-steekproeven uit de 10-fold cross-validation is terug te vinden in tabel 5. Op het eerste gezicht lijkt er geen opmerkelijk verschil tussen de versie waarbij TreeTalk op de hoogte is van de samenstellingen en de klassieke TreeTalk. Statistische toetsing bevestigt dit vermoeden: er is geen significant verschil tussen het gemiddelde van beide varianten[13]. Ook niet als het outputfilter gebruikt wordt[14].

3.3.2 Phone Error Rate

Een exactere maat van correctheid bij de grafeem-foneemconversie is de Phone Error Rate. Naar analogie met de WER staan de resultaten voor de PER in tabel 6 . Opnieuw is er geen significant verschil als TreeTalk over morfologische features beschikt[15], ook niet als nadien het outputfilter toegepast wordt[16]

3.3.3 Verschillende fouten op woordniveau

Zoals te zien is in tabel 7 hebben TreeTalk en MTreeTalk problemen met grotendeels dezelfde woorden.

[13] $p = 0.889$; alle statistische analyses zijn te vinden in bijlage A
[14] $p = 0.831$
[15] $p = 0.336$
[16] $p = 0.256$

Tabel 5: WER met en zonder morfologische informatie.

Testset	**TreeTalk**		**MTreeTalk**	
	Zonder filter	*Met filter*	*Zonder filter*	*Met filter*
1	17.24%	17.12%	17.28%	17.15%
2	17.84%	17.73%	17.55%	17.46%
3	17.32%	17.20%	17.28%	17.16%
4	17.19%	17.01%	17.20%	17.02%
5	17.45%	17.32%	17.41%	17.29%
6	17.57%	17.44%	17.78%	17.63%
7	17.57%	17.45%	17.59%	17.48%
8	16.95%	16.80%	16.98%	16.84%
9	17.58%	17.48%	17.50%	17.39%
10	17.51%	17.37%	17.72%	17.58%
Gemiddelde	**17.42%**	**17.29%**	**17.43%**	**17.30%**

Tabel 6: PER met en zonder morfologische informatie.

Testset	**TreeTalk**		**MTreeTalk**	
	Zonder filter	*Met filter*	*Zonder filter*	*Met filter*
1	3.06%	2.02%	3.09%	2.03%
2	3.23%	2.11%	3.17%	2.05%
3	3.11%	2.03%	3.11%	2.01%
4	3.10%	1.98%	3.12%	1.98%
5	3.12%	2.03%	3.12%	2.02%
6	3.15%	2.04%	3.21%	2.05%
7	3.14%	2.05%	3.15%	2.05%
8	3.05%	1.95%	3.04%	1.95%
9	3.17%	2.02%	3.18%	2.01%
10	3.14%	2.03%	3.18%	2.05%
Gemiddelde	**3.13%**	**2.03%**	**3.14%**	**2.02%**

Als de fonemisatie van een woord een fout bevat bij de ene, zal dat in meer dan 90% van de gevallen ook bij de andere versie zo zijn.

Tabel 7: Percentage unieke fouten

Testset	Zonder filter	Met filter
1	9.29%	9.42%
2	9.45%	9.59%
3	9.89%	9.99%
4	9.23%	9.35%
5	9.15%	9.28%
6	9.55%	9.68%
7	9.65%	9.76%
8	10.01%	10.14%
9	9.40%	9.52%
10	9.59%	9.73%
Gemiddelde	**9.52%**	**9.64%**

4 Discussie

4.1 Kwantitatieve analyse

Allereerst kunnen we vaststellen dat de prestaties van TreeTalk ondanks de extra informatie tijdens training en classificatie niet verbeteren. Dit kan verschillende zaken betekenen:

- Morfologische informatie over samenstellingen is bij inductieve grafeem-foneem-omzetting:
  - Van geen of ondergeschikt belang.
  - Van belang, maar heeft al grotendeels dezelfde invloed als de grafemische informatie. De morfologie zit als het ware al impliciet ingebakken in de lettercontext van het focusfeature.
- De samenstellingsinformatie waarover TreeTalk in dit experiment beschikte is te beperkt (zie sectie 4.1.2) of bevat te veel fouten.

4.1.1 Invloed van de gebruikte features

Om dieper in te gaan op de eerste veronderstelling is het interessant om te kijken naar het relatieve belang van de grafeem- en samenstellingsfeatures. Hoe groot is de invloed van de morfologische features op de classificatie? Een nadere blik op de "Information Gain Ration" (IGR) leert dat deze niet te verwaarlozen is. Deze maat geeft aan hoeveel gewicht een bepaald feature gekregen heeft tijdens de trainingsprocedure,

met andere woorden: hoe descriptief is een bepaald feature? Het gaat om een genormaliseerde maat die de invloed van het aantal klasses tot op een bepaalde hoogte neutraliseert[17].

In figuur 3 staat een overzicht van de IGR voor alle features die bij TreeTalk gebruikt werden. F5 is het focusfeature, F1 tot F4 vormen de linkercontext terwijl F6 tot F9 het belang van de rechtercontext aangeven. Er blijkt duidelijk dat het focusfeature het belangrijkste is, gevolgd door respectievelijk de onmiddellijke rechter- en linkerbuur. Dit is volledig in lijn met [Daelemans and van den Bosch, 1993].

Figuur 4 geeft vervolgens hetzelfde overzicht voor MTreeTalk, met dat verschil dat nu ook het belang van de morfologische features opgenomen is. F12 is daarbij het focusfeature voor de samenstellingsanalyse, F10 en F11 vormen de linkercontext, F13 en F14 de rechtercontext. Het relatieve belang van de samenstellingsinformatie wordt hierdoor ook aangetoond. Na het focusgrafeem (met linker- en rechterbuur) volgt de samenstellingsbit van het morfologische focusfeature (eveneens met linker- en rechterbuur). Hoewel –logischerwijze– de focusletter de grootste invloed blijft hebben kunnen we toch gerust stellen dat ook de morfologie in rekening wordt gebracht bij de classificatie.

4.1.2 Mogelijke verklaringen

Uit voorgaande gegevens blijkt dat MTreeTalk gebruik maakt van de morfologische informatie maar dat dit niet tot een prestatieverbetering leidt. Meer nog, ook de gemaakte fouten blijven grotendeels dezelfde. Hier volgen twee mogelijke verklaringen.

Allereerst zou het kunnen dat de morfologische analyses die voor de training gebruikt werden regelmatig foutief waren. Uit praktische overwegingen hebben we immers MBMA gebruikt als basis voor de morfologische analyses en niet een (gegarandeerd correcte) analyse uit een corpus als CELEX. Om dit de falsificeren zou het interessant zijn om hetzelfde experiment te repliceren met “zuivere” morfologische gegevens voor de trainingsdata.

Als tweede reden kunnen we denken aan het feit dat de morfologische informatie vaak impliciet opgeslagen kan zijn in de grafemische weergave van een woord. Als TreeTalk “petje” aan het einde van een woord als /p@tj@/ uitsrpreekt, dan geeft het daarmee te kennen dat het zich ervan “bewust” is dat hier om een diminutiefvorm gaat die met een sjwa uitgesproken wordt. Mogelijk zijn er weinig gevallen waarbij de expliciete morfologie nodig is of volstaat om tot een betere fonemisatie te komen.

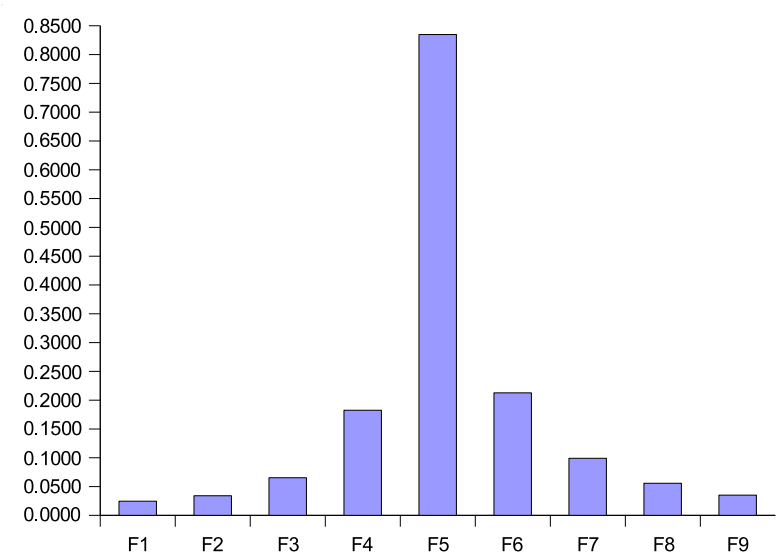
4.2 Kwalitatieve analyse

Waar liggen precies de verschillen tussen TreeTalk en MTreeTalk? Om dit te illustreren geven we in tabel 8 enkele voorbeelden van uitvoer die bij beide versies verschilt. Het is een willekeurige selectie die niet pretendeert een volledig overzicht te geven. Niettemin kan het een idee geven van de onderlinge verschillen in de praktijk.

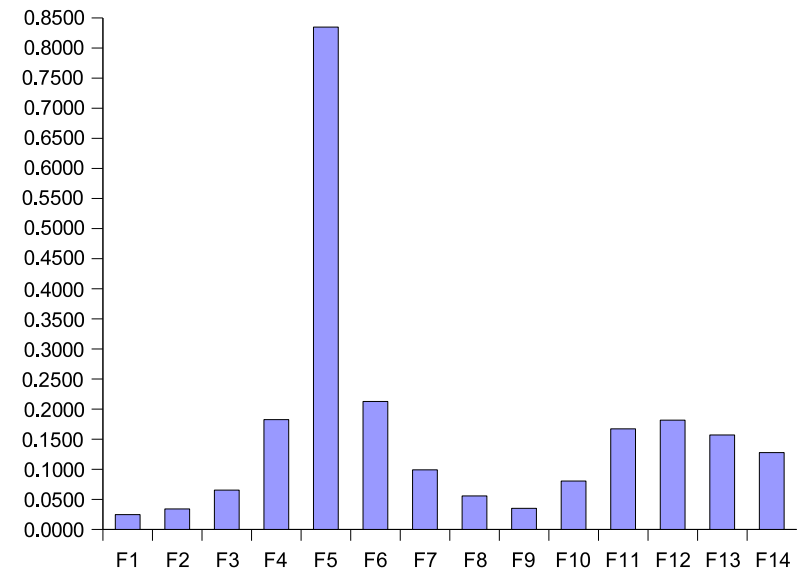
Enkele vaststellingen hierbij:

- De tussen-n komt bij beide varianten soms wel en soms niet voor (vragendag, boekenstut).

[17] De niet-genormaliseerde maatstaaf, de Information Gain, heeft de neiging om features waarvoor veel verschillende klasses bestaan als belangrijker te markeren. Voor een uitgebreide discussie van deze maten verwijzen we naar [Daelemans et al., 2004].



Figuur 3: Information Gain Ratio voor TreeTalk



Figuur 4: Information Gain Ratio voor MTreeTalk

- Klinkers vallen soms weg op bepaalde plaatsen, in beide gevallen (aardnoot, vechtgraag).
- Verschillende fouten zijn nauwelijks of niet hoorbaar (dronkelappen, zinvolle, godsbegrip).
- Sommige hoogfrequente woorden gaan totaal mis (wou, meegaan) maar dit zal bij praktijkgebruik niet hinderlijk zijn aangezien net die woorden in een vast uitspraaklexicon opgeslagen zijn.
- Er is geen eenduidig onderscheid te maken tussen fouten die TreeTalk en MTree-Talk maken.

Tabel 8: Voorbeelden van verschillende resultaten bij TreeTalk en MTreeTalk (zonder outputfilter)

Woord	TreeTalk	MTreeTalk
vragendag	((((v r a) 1) ((G @) 0)) (((d A x) 2)))	((((v r a) 1) ((G @ n) 0)) (((d A x) 2)))
boekenstut	((((b u) 1) ((k @ n) 0)) (((s t Y t) 2)))	((((b u) 1) ((k @) 0)) (((s t Y t) 2)))
aardnoot	((((a r d n t) 1)))	((((a r d) 1)) (((n o t) 2)))
vechtgraag	((((v E x t) 1)) (((x r a x) 2)))	((((v E x t) 1)) (((x r x) 2)))
clublid	((((k l y) 1) ((p l I t) 0)))	((((k l y b) 1)) (((l I t) 2)))
voorst	((((v o r s t) 1)))	((((v o) 1) ((o r s t) 0)))
migratie-geschiedenis	((((m i) 0) ((G r a) 1) ((t s i) 0)) (((x @) 0) ((s x i) 2) ((d @) 0) ((n I s) 0)))	((((m i) 0) ((G r a) 1) ((t s i) 0)) (((G @) 0) ((s x i) 2) ((d @) 0) ((n I s) 0)))
dronkelappen	((((d r O N) 1) ((k @) 0)) (((l A) 2) ((p @ n) 0)))	((((d r O N) 1) ((k @) 0)) (((l A p) 2) ((p @ n) 0)))
zinvolle	((((z I n) 1)) (((v O) 2) ((l @) 0)))	((((z I n) 1)) (((v O l) 2) ((l @) 0)))
wou	((((w A+ w) 1)))	((((w) 1)))
meegaan	((((m) 0 -)) (((G a n) 2)))	((((m G n) 0)))
godsbegrip	((((x O d z) 1)) (((b @) 0) ((G r I p) 2)))	(((x O d z) 1)) (((b @) 0) ((x r I p) 2)))
puntlassen	((((p Y n t) 1)) (((l A) 2) ((s @ n) 0)))	(((p Y n t) 1)) (((l) 2 -) ((A @ n) 0 +)))

Een woord waarbij de potentiële sterkte van MTreeTalk naar boven komt is "morgenster" (/mOrG@stEr/). In [Daelemans and van den Bosch, 1993] wordt al aangegeven dat TreeTalk een linkercontext van 5 letters nodig heeft om de laatste "e" correct in een /E/ om te zetten. Zoniet dan weegt de analogie met onder andere "venster" en "dienster" te zwaar door, waarbij de laatste "e" telkens als /@/ uitgesproken wordt. MTreeTalk geeft ook met slechts een linkercontext van 4 grafemen de goede fonemisatie. Vermoedelijk kan in dit geval het onderscheid tussen samenstellingen en niet-samenstellingen optimaal benut worden.

Al bij al is het echter lastig om de precieze aard weer te geven van de verschillen tussen TreeTalk en MTreeTalk. Niet alleen zijn veel fouten gelijkaardig, ook de

gebruikte inductieve leermethode laat alleen een interpretatie toe die geval per geval bekeken moet worden. Daartoe ontbraken de tijd en middelen.

4.3 Voorstellen voor vervolgonderzoek

Vanzelfsprekend is binnen het beperkte kader van dit onderzoek niet alles onderzocht wat interessant leek. Daarom volgen hier nog enkele ideeën en suggesties voor toekomstig onderzoek.

4.3.1 Het gebruik van andere TiMBL-parameters

Zoals aangegeven in sectie 2.1.1 maken alle genoemde inductieve leersystemen (i.e. TreeTalk en MBMA, beiden aangedreven door TiMBL) gebruik van het IGTree-algoritme. Dit werkt sneller en gebruikt minder geheugen dan het standaardalgoritme (IB1) maar presteert dus ook suboptimaal. Vanuit een praktisch standpunt is dit te verkiezen; voor onderzoeksdoeleinden zou het echter interessant zijn om het experiment ook eens uit te voeren met IB1.

Daarnaast zouden ook de parameters van TiMBL geoptimaliseerd kunnen worden. Paramsearch[18] is een programma dat hiervoor speciaal ontworpen is. Het zoekt iteratief naar de instellingen waarmee de beste classificatie bereikt wordt, door de zoekruimte in de parameterruimte keer op keer te verkleinen.

4.3.2 Perfecte samenstellingsgrenzen

Bij de training van MTreeTalk werd gebruik gemaakt van morfologische informatie zoals die door MBMA gegenereerd werd. We hebben het dus over een benadering van de samenstellingsgrenzen (zoals uiteengezet in sectie 4.1.2). Als alternatief zou MTreeTalk getraind kunnen worden met perfecte samenstellingsgrenzen, zoals die bijvoorbeeld te vinden zijn in CELEX. De hamvraag is in dat geval of de hogere precisie qua morfologie ook voor een merkbare verbetering zou zorgen op het gebied van de grafeem-foneemconversie.

4.3.3 Een breder venster voor de morfologische features

Uit figuur 4 bleek al dat de morfologische features een zekere invloed hebben bij de classificatie bij MTreeTalk. Het zou interessant te zijn om te onderzoeken wat de optimale breedte is voor het venster met morfologische features.

4.3.4 Aparte fonemisatie voor samenstellingsleden

Het gegeven dat een woord een samenstelling is kan ook op een andere manier aangewend worden dan nu het geval is bij MTreeTalk. In plaats van de fonemisatie voor het hele woord uit te voeren zouden we eerst de samenstelling kunnen opsplitsen in de verschillende onderdelen. Vervolgens kan een grafeem-foneemconversie toegepast worden op elk van de delen. Tot slot kunnen deze uitgesproken samenstellingsleden weer aan elkaar gekoppeld worden. Schematisch zou het er bijvoorbeeld als volgt uitzien:

[18] `http://ilk.uvt.nl/mblp/`

```
"kamelenrace"
> MBMA: "kamelen" + "race"
> Grafeem-foneemconversie: /kamel@/ + /res/
> Aaneenschakeling: /kamel@res/
```

Op die manier zou de tekst-naar-spraakomzetting van samenstellingen als “kamelenrace” kunnen profiteren van de kennis die opgeslagen is in het uitspraaklexicon. “Kamelen” en “race” komen er namelijk in voor, “kamelenrace” niet, met als gevolg dat TreeTalk en MTreeTalk het foutieve /kam@l@res/ als uitvoer genereren. Hetzelfde geldt voor de inductieve grafeem-foneemconversie: die gaat vaak beter bij de samenstellingsleden aangezien er in die gevallen geen misleiding kan ontstaan door een vreemde context.

Aanvullend bij deze techniek zouden eventueel de resultaten van [Wuite, 2004] gebruikt kunnen worden om na te gaan of de detectie van een nominale samenstelling al dan niet terecht is. Uit dat onderzoek blijkt namelijk dat de lengte van de constituenten een goede heuristiek leveren voor de onterechte “herkenning” van nominale samenstellingen[19]. Als alle constituenten langer zijn dan 3 karakters kunnen we er van uit gaan dat het wel degelijk om een samenstelling van substantieven gaat. In het andere geval is de kans op een vals-positieve herkenning zeer groot.

Als bijkomend voordeel zouden de samenstellingsconstituenten ook doorgestuurd kunnen worden naar FonPars in plaats van TreeTalk. Eerstgenoemde ondersteunt namelijk geen samenstellingen; een aparte analyse van de samenstelling zou dit gebrek dus oplossen.

Uiteraard zijn er ook aan deze methode nadelen verbonden. Samenstellingen in het Nederlands kennen vaak een vrij onregelmatig en onvoorspelbaar patroon, al was het maar omwille van de bindingsfonemen “en”, “e” en “s”. Tekenend voor de complexiteit hiervan zijn de recente discussies over de spellingshervorming voor het Nederlands. Om nog te zwijgen over fenomenen als medeklinkerverdubbeling (“muggen” is geen samenstelling van “mug” en “gen”).

5 Conclusie

5.1 Antwoord op de onderzoeksvraag

De algemene slotconclusie waarmee we dit onderzoek kunnen afsluiten is dat het toevoegen van samenstellingsinformatie aan de inductieve grafeem-foneemconversie van TreeTalk geen significante invloed heeft.

5.2 Hypotheses

We overlopen nog eens de hypotheses zoals geformuleerd in sectie 2.3.2.

5.2.1 Hypothese 1

> Inductieve grafeem-foneemomzetting presteert even goed mét als zónder extra morfologische informatie tijdens de training en classificatie.

[19] “Oorlog” kan bijvoorbeeld ten onrechte gezien worden als de samenstelling van de nomina “oor” en “log”.

Deze hypothese is correct. Er is geen significant verschil te vinden tussen TreeTalk en MTreeTalk.

5.2.2 Hypothese 2

> De grafeem-foneemconversie met morfologie maakt dezelfde fouten als die zonder morfologie.

Ook deze stelling blijkt waar te zijn. Minder dan 10% van de gemaakte fouten is uniek op woordniveau. In alle andere gevallen komen bij beide varianten de fouten voor in dezelfde woorden.

5.3 Fouten: aantal en aard

De Word Error Rate is voor TreeTalk gemiddeld 17.42%, voor MTreeTalk 17.43%. Met een outputfilter wordt dat respectievelijk 17.29% en 17.30%.

De Phone Error Rate bij TreeTalk bedraagt 3.13%, bij MTreeTalk 3.14%. Als de resultaten ook door een outputfilter verwerkt worden komen de scores op respectievelijk 2.03% en 2.02%.

Referenties

[Baayen et al., 1993] Baayen, R., Piepenbrock, R., and van Rijn, H. (1993). The CELEX lexical database (CDROM). `http://www.ru.nl/celex/subsecs/section_doc.html`.

[Barton et al., 1987] Barton, G. E., Berwick, R., and Eric Ristad, E. S. (1987). *Computational Complexity and Natural Language*. MIT Press.

[Booij and van Santen, 1998] Booij, G. and van Santen, A. (1998). *Morfologie: de woorstructuur van het Nederlands*. Amsterdam University Press.

[Busser, 1998] Busser, B. (1998). TreeTalk-D: a machine learning approach to Dutch word pronunciation. Proceedings of the TSD Conference, pages 3–8.

[Chomsky and Halle, 1968] Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Harper & Row.

[Daelemans and van den Bosch, 1993] Daelemans, W. and van den Bosch, A. (1993). Data-oriented methods for grapheme-to-phoneme conversion. In *Proceedings of the sixth conference of the European chapter of the ACL*, pages 45–53.

[Daelemans et al., 2004] Daelemans, W., Zavrel, J., van der sloot, K., and van den Bosch, A. (2004). TiMBL Reference Guide version 5.1. Technical report, ILK.

[Gillis, 2001] Gillis, S. (2001). Protocol voor brede fonetische transcriptie van het CGN. `http://lands.let.kun.nl/cgn/doc_Dutch/topics/version_1.0/annot/phonetic%s/info.htm#protocol`.

[Kerkhoff and Rietveld, 1994] Kerkhoff, J. and Rietveld, T. (1994). Prosody in Niros with Fonpars and Alfeios. volume 18 of *Proceedings of the Dept. of Language & Speech, University of Nijmegen*, pages 107–119.

[Kerkhoff et al., 1998] Kerkhoff, J., Rietveld, T., and van Bergem, D. (1998). Het Nijmeegse tekst-naar-spraaksysteem KunTTS. Department of language and speech, University of Nijmegen.

[Konings, 2003] Konings, N. (2003). Leren of laten leren? een vergelijking van regelgebaseerde en zelflerende grafeem-foneemomzetting voor het Nederlands. Master's thesis, Universiteit van Tilburg.

[Marsi and Kerkhoff, 2002] Marsi, E. and Kerkhoff, J. (2002). NeXTeNS, A new open source text-to-speech system for Dutch. `http://nextens.uvt.nl/presentations/clin02-nextens.htm`.

[Marsi and Kerkhoff, 2003] Marsi, E. and Kerkhoff, J. (2003). NeXTeNS website. `http://nextens.uvt.nl/`.

[Sejnowski and Rosenberg, 1987] Sejnowski, T. J. and Rosenberg, C. (1987). Paralel networks that learn to pronounce English text. *Complex Systems*, (1):145–168.

[Tarsaku et al., 2001] Tarsaku, P., Sornlertlamvanich, V., and Thongprasirt, R. (2001). Thai Grapheme-to-Phoneme Using Probabilistic GLR Parser. volume 2 of *Proceedings of Eurospeech*, pages 1057–1060.

[van den Bosch and Daelemans, 1999] van den Bosch, A. and Daelemans, W. (1999). Memory-based morphological analysis. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 285–292, Morristown, NJ, USA. Association for Computational Linguistics.

[van den Bosch and Daelemans, 2005] van den Bosch, A. and Daelemans, W. (2005). *Memory-based language processing*. Cambridge University Press.

[Wagner and Fischer, 1974] Wagner, R. A. and Fischer, M. J. (1974). The string-to-string correction problem. *J. ACM*, 21(1):168–173.

[Weiss and Kulikowski, 1991] Weiss, S. and Kulikowski, C. (1991). *Computer systems that learn*. Morgan Kaufmann.

[Wuite, 2004] Wuite, M. (2004). Compound recognition and its application in spelling correction. Master's thesis, University of Nijmegen.

A Statistische analyses

A.1 Word Error Rate

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	TreeTalk	.1742	10	.00251	.00079
	MTreeTalk	.1743	10	.00249	.00079
Pair 2	TreeTalkFilter	.1729	10	.00269	.00085
	MTreeTalkFilter	.1730	10	.00257	.00081

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	TreeTalk & MTreeTalk	10	.841	.002
Pair 2	TreeTalkFilter & MTreeTalkFilter	10	.862	.001

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	TreeTalk - MTreeTalk	-.00006	.00141	.00045	-.00107	.00095	-.144	9	.889
Pair 2	TreeTalkFilter - MTreeTalkFilter	-.00010	.00138	.00044	-.00109	.00089	-.220	9	.831

A.2 Phone Error Rate

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	TreeTalk	.031258	10	.0005348	.0001691
	MTreeTalk	.031368	10	.0005102	.0001614
Pair 2	TreetalkFilter	.020263	10	.0004042	.0001278
	MTreeTalkFilter	.020189	10	.0003491	.0001104

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	TreeTalk & MTreeTalk	10	.785	.007
Pair 2	TreetalkFilter & MTreeTalkFilter	10	.878	.001

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	TreeTalk - MTreeTalk	-.0001102	.0003432	.0001085	-.0003557	.0001353	-1.015	9	.336
Pair 2	TreetalkFilter - MTreeTalkFilter	.0000742	.0001935	.0000612	-.0000642	.0002127	1.213	9	.256

B Foneemset CGN

Bron: [Konings, 2003]

Vocalen	Voorbeeld		Consonanten	Voorbeeld
@	de		p	pas
A	pad		t	tas
E	pet		k	kas
I	pit		b	bas
O	pot		d	das
Y	put		g	goal
i	vier		f	fiets
u	voer		v	vaas
y	vuur		s	sap
a	laan		z	zeep
e	veer		S	sjiek
o	rood		Z	gage
2	deur		x	toch
E+	reis		G	regen
Y+	huis		h	hand
A+	koud		m	man
E:	beige		n	nam
Y:	freule		N	lang
O:	roze		J	oranje
E~	vaccin		r	rand
A~	croissant		l	lief
O~	congé		j	jas
Y~	parfum		w	wat