MA Taalkunde: Masterproef-presentatie

Abstract

Naam:	Adriaan **Lemmens**
Titel:	*DePicto Dutch 0.1: an open-source HPSG-based pictograph-to-text translation framework for use in alternative communication applications*
Promotor:	Vincent Vandeghinste
Beoogde voltooiing:	September

Opdracht

Een *proof-of-concept* **systeem** ontwikkelen dat aan de hand van een **regel-gebaseerd taalmodel** een reeks ingevoerde **pictogrammen** in een of meer welgevormde zinnen vertaalt. De doeltaal in de huidige toepassing is het **Nederlands**, hoewel de kern van het systeem in principe taal-onafhankelijk is. Het doelpubliek bestaat voornamelijk uit gebruikers die, wegens een **intellectuele beperking** en/of ontwikkelingsstoornis, voorlopig geen, of slechts beperkt, gebruik kunnen maken van online communicatiemiddelen.

Achtergrond en probleemstelling

Het vermogen om sociaal actief te zijn in de digitale wereld is een factor die steeds meer bepalend wordt voor de levenskwaliteit. Doordat de meeste activiteiten echter een schriftelijke basis hebben, is zelfstandige participatie voor mensen met intellectuele beperkingen meestal uitgesloten. Om deze drempel te verlagen moet er een alternatief schriftelijk communicatiemiddel aangeboden kunnen worden dat door zijn simpliciteit maximaal toegankelijk is. Een goede kandidaat in dit verband is een pictografisch 'schrift', dat binnen de Europese Unie alleen al twee tot vijf miljoen mensen zou kunnen helpen (Keskinen et al., 2012) bij het lezen en schrijven van natuurlijke taal.

Software implementaties die voor de vertaling van natuurlijke taal naar pictografische systemen zorgen, bestaan al in redelijk goed uitgewerkte en gebruikersvriendelijke vorm (zie o.a. Vandeghinste et al. 2015). De omgekeerde richting daarentegen, nl. van pictogrammen naar tekst, heeft dusver veel minder aandacht gekregen, vooral met betrekking tot gebruikersvriendelijkheid. De uitzondering in dit verband is Sevens et al. (2015), dat een pictogram-naar-tekst vertaalsysteem voorstelt waarin zo weinig mogelijk 'structurele' taalkennis van de gebruiker wordt verwacht, maar aan een statistisch taalmodel wordt overgelaten. Het systeem heeft veel potentieel, maar heeft ook moeite met bepaalde eigenaardigheden van de gekozen pictogramtaal. (Bijvoorbeeld: 'naamwoord'-pictogrammen

drukken geen onderscheid tussen meervoud en enkelvoud uit; 'werkwoord'-pictogrammen drukken geen aspect of tijd uit; hulpwerkwoorden en lidwoorden hebben geen pictografische pendant (Sevens et al., 2015); pictogrammen kunnen complexe situaties uitdrukken, waardoor er geen sprake van een een-op-een relatie tegenover de doeltaal kan zijn.) In tegenstelling tot een statistisch taalmodel, zou een regel-gebaseerd (d.i. formeel) taalmodel dergelijke kenmerken relatief moeiteloos moeten kunnen incorporeren, mits voldaan wordt aan de eisen van het gekozen formalisme.

Om die bewering te testen – en hopelijk bevestigen – wordt in het kader van deze masterproef het prototypesysteem *DePicto (Dutch)* ontwikkeld. Het systeem neemt een deel van de functionaliteit van het systeem voorgesteld in Sevens et al. (2015) over, maar doet dat vanuit een louter regel-gebaseerde benadering (hierover zometeen meer). Daarbij worden in de eerste plaats syntactische hoofdzinfenomenen (congruentie, determinatie, reflexiviteit, tijd en aspect, eventueel modus-verandering) en in de tweede plaats – indien er tijd overblijft – complexe pictogram-predicaten behandeld. Daarnaast gaat er steeds aandacht naar de vraag of het praktisch haalbaar zou zijn om het systeem volledig uit te werken, alsook uit te breiden naar andere talen. Overige aandacht gaat naar het produceren van documentatie dat toekomstig onderhoud zal vergemakkelijken.

Het systeem

De structuur van het *DePicto*-systeem is stabiel. Het is gebaseerd rond twee grammatica's geschreven in het HPSG-formalisme[1] (Sag et al. 2003, Pollard & Sag 1994) en ontwikkeld en gecompileerd in het *LKB*-systeem[2] (Copestake, 2002). De eerste van deze twee grammatica's krijgt als invoer een reeks van lemma-IDs geassocieerd met de ingevoerde pictogrammen en probeert die te ontleden om tot een welgevormde structuur te komen. Indien zo'n structuur gevonden wordt, levert deze eerste grammatica een semantische representatie (Copestake, 2005) van de ingevoerde 'pictogram-zin' op. De semantische representatie kan elementen bevatten die op de inherent ondergespecificeerde invoer niet terug te vinden zijn. Zo worden ontbrekende lidwoorden bijvoorbeeld toegevoegd door middel van een kleine familie van *phrase rules* die eigen zijn aan deze grammatica. De tweede grammatica neemt als invoer de semantische representatie bekomen in de vorige stap en genereert op basis hiervan alle hypothetisch mogelijke welgevormde zinnen. Omdat de doeltaal in de huidige toepassing het Nederlands is, is de grammatica hierop ingesteld. Maar de kern van de grammatica, die ontleend is aan de open-source LinGO Grammar Matrix (Bender 2002), is grotendeels taal-onafhankelijk. Alle 'regels' die specifiek aan het Nederlands zijn, worden in een apart bestand bewaard, en vullen deze algemene kern enkel aan. De eerste grammatica heeft overigens dezelfde kern, waardoor de onderliggende architectuur van beide grammatica's identiek hetzelfde is.

[1] HPSG (Head-Driven Phrase Structure Grammar) behoort tot de familie van unificatie-gebaseerde formalismen. Het bewaart grammaticale informatie in getypeerde kenmerkstructuren, gebruikt de operatie van unificatie om die kenmerkstructuren met elkaar te combineren tot 'grotere' structuren (waaruit onder andere constituentenbomen kunnen worden afgeleid), en plaatst de nadruk vooral op het lexicon.

2 Latere versies zullen gebruik maken van het efficiëntere *ACE*-systeem (Woodley, 2015).

Stand van zaken

Afgeronde taken

- Formalisme kiezen en grondig leren gebruiken.
- Basisarchitectuur vastleggen en methodes verzinnen om de automatische invoer-uitvoerstroom tussen twee grammatica's te bewerkstelligen.
- Een eerste prototype van een Nederlandse grammatica ontwikkelen (op basis van de geërfde kern); testen dat die welgevormde Nederlandse zinnen kan genereren.
- Een voorlopige grammatica ontwikkelen (o.b.v. bovenvermelde prototype) dat simpele 'pictogram-zinnen' kan ontleden, mits die zich aan een SVO(O) volgorde houden.
- Mechanisme ontwikkelen binnen het HPSG-formalisme dat nominale constituenten met een zelfstandignaamwoordhoofd op het semantisch niveau van een lidwoord voorziet.

Wat er nog moet gebeuren

- De dekking van de pictogram-grammatica uitbreiden m.b.t. (een selectie (waarschijnlijk) van) de volgende fenomenen: negatie, ja-nee-vragen, co-ordinatie, tijd (misschien aspect), en pictogrammen die complexe situaties uitdrukken.
- Nederlandse V2 volgorde implementeren (moeilijk want meerdere analyses mogelijk).
- Documentatie.
- De prestaties van het systeem vergelijken met dat voorgesteld in Sevens et al. (2015).
- Een scriptie schrijven.

Open vragen/Problemen

- Momenteel wordt er impliciet van uitgegaan dat gebruikers zich de SVO-woordvolgorde eigen hebben kunnen maken (andere volgordes uiteraard mogelijk voor andere talen). Dat zou een foute aanname kunnen zijn. Mocht dat het geval blijken te zijn zou het systeem voorzien moeten worden van een mechanisme dat vrije woordvolgorde toestaat. Daaruit ontstaat echter een nieuw probleem: zonder een vaste/voorspelbare woordvolgorde (waarbij ook geen naamvalmarkering komt kijken), kunnen allerlei soorten ambiguïteit ontstaan: o.a. het onderscheid tussen onderwerp en voorwerp zou verloren gaan.
- Doordat de huidige grammatica's vanuit een redelijke vaste 1-op-1 verhouding gaan tussen elementen in het lexicon en pictogrammen is er geen sprake van synonymie. Patronen zoals collocatie worden hierdoor moeilijk te dekken. Ook gevoelswaarde en idioom komen in het gedrang.

Bibliografie

Bender, E.M., Flickinger, D., Oepen, S. 2002. The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*. Taipei, Taiwan. pp. 8-14

Bender, E.M. 2001. *Syntactic Variation and Linguistic Competence: The Case of AAVE Copula Absence*. PhD thesis, Stanford University.

Bender, E.M. et al. 2010. Grammar Prototyping and Testing with the LinGO Grammar Matrix Customization System. In *Proceedings of the ACL 2010 System Demonstrations.* Sweden. 1–6.

Bender, E.M. 2013. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax.* Synthesis Lectures on Human Language Technologies 6(3).

Bouma, G. & van Noord, G. 1996. Word Order Constraints on Verb Clusters in German and Dutch. In *Proceedings of Formal Grammar 96*.

Bouma, G., Van Eynde, F. & Flickinger, D. 1998. Constraint-based Lexicons. In Van Eynde, F. & Gibbon, D.(eds.) *Lexicon Development for Speech and Language Processing*. 43–75. Dordrecht: Kluwer.

Carroll, J., Copestake, A., Flickinger, D. & Poznanski, V. 1999. An Efficient Chart Generator for (Semi-)Lexicalist Grammars. In *Proceedings of the 7th European Workshop on Natural Language Generation* (EWNLG'99). Toulouse.

Copestake, A. 2002. *Implementing Typed Feature Structures.* CSLI Publications, 2002.

Copestake, A. & Flickinger, D. 2000. An open source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the LREC*. 591–600.

Copestake, A., Lascarides, A. & Flickinger, D. 2001. An Algebra for Semantic Construction in Constraint-Based Grammars. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (ACL 2001), Toulouse, France.

Copestake, A., Flickinger, D., Pollard, C. & Sag, I.A. 2005. "Minimal Recursion Semantics: An Introduction." *Journal of Research on Language and Computation*, 3(2-3): 281-332.

Crysmann, B. & Packard W. 2012. Towards efficient HPSG generation for German, a non-configurational language. In: *Proceedings of COLING 2012: Technical Papers.* 695–710.

Flickinger, D. & Bender, E.M. 2003. Compositional Semantics in a Multilingual Grammar Resource. In *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development*, ESSLLI. 33–42.

Flickinger, D., Bender, E.M., Oepen, S. 2003. MRS in the LinGO Grammar Matrix: A Practical User's guide. http://faculty.washington.edu/ebender/papers/userguide.pdf (March 2016)

Fokkens, A. 2011. Metagrammar engineering: Towards systematic exploration of implemented grammars. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 1066–1076. Portland, Oregon, USA: Association for Computational Linguistics.

Keskinen, T., Heimonen, T., Turunen M., Rajaniemi, J.P. & Kauppinen, S. 2012. SymbolChat: A Flexible Picture-based Communication Platform for Users with Intellectual Disabilities. *Interacting with Computers*, 24(5): 374–386.

Krieger, H. & Schäfer, U. 1994. "TDL—A Type Description Language for Constraint-Based Grammars." In *Proceedings of the 15th International Conference on Computational Linguistics*, COLING-94. 893–899.

Levine, R.D., Meurers, W.D. 2006. Head-Driven Phrase Structure Grammar: Linguistic Approach, Formal Foundations, and Computational Realization. In: Keith Brown (Ed.): *Encyclopedia of Language and Linguistics,*

Second Edition. Oxford: Elsevier. 2006.

Müller, S. 2013. HPSG – A Synopsis. In: Alexiadou, A., Kiss, T. (eds.) *Syntax – Ein internationales Handbuch zeitgenössischer Forschung, 2nd edition*. Walter de Gruyter Verlag, Berlin [to appear].

Pollard, C.J. & Sag, I.A. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Sag, I.A., Wasow, T. & Bender, E.M. 2003. *Syntactic Theory: A formal introduction, Second Edition*. Stanford: CSLI Publications [distributed by University of Chicago Press].

Sevens, L., Vandeghinste, V., Schuurman, I. & Van Eynde, F. 2014. "Improving the Precision of Synset Links Between Cornetto and Princeton WordNet." In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*. Coling 2014. Dublin, Ireland.

Sevens, L., Vandeghinste, V., Van Eynde, F. 2015. "Natural Language Generation from Pictographs." In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG).* Association for Computational Linguistics. 71-75.

Vandeghinste, V. & Schuurman, I. 2014. "Linking Pictographs to Synsets: Sclera2Cornetto." LREC 2014. Reykjavik, Iceland.

Vandeghinste, V., Schuurman, I., Sevens, L. & Van Eynde, F. 2015. Translating Text into Pictographs. Natural Language Engineering. http://picto.ccl.kuleuven.be/publications.html#sthash.zBS5flcW.dpuf