20 minute presentation slot

Multiword proper names in compounds and derivations: Towards a morphographemic model of choice

Gerhard van Huyssteen

Centre for Text Technology (CTexT), North-West University, Potchefstroom

gerhard.vanhuyssteen@nwu.ac.za

This research deals with rather peripheral constructions in Afrikaans morphology, viz. compounds with and derivations of multiword proper names (MWPs), such as:

(1) MWP: *Middellandse See* ('Mediterranean Sea')
Compound: *Middellandse See* + *gebied* ('area')
Derivation: *Middellandse See*·ADJZ ('Mediterranean Sea-ADJZ')
(2) MWP: *Dooie See* ('Dead Sea')
Compound: *Dooie See* + *rolle* ('scrolls')
Derivation: *Dooie See*·ADJZ

We are specifically interested in the orthographic realisation of such complex forms, and as such the research should be seen as an investigation into the interface between morphology and orthography – also sometimes referred to as morphographemics, i.e. "… the area dealing with systematic discrepancies between the surface form of words and the symbolic representation of the words in a lexicon. Such differences are typically orthographic changes that occur when basic lexical items are concatenated …" (Black et al. 1987).

If we consider a representation such as [[*Middellandse See*]$_{MWPi}$ [*gebied*]$_{Nj}$]$_{Nk}$ ⇔ [SEM$_j$ of SEM$_i$]$_k$ as the symbolic representation of the specific compound, the question is how this construction is realised on the orthographic pole (i.e. the pole of realisation, conventionally called the phonological pole in Cognitive Grammar). For example, it could be realised potentially as *Middellandse Seegebied; Middellandse See-gebied; Middellandseseegebied; Middellandse-Seegebied*; etc. Similarly, the construction [[*Middellandse See*]$_{MWPi}$ [*s*]$_{ADJZ}$]$_{Nk}$ ⇔ [related to SEM$_i$]$_k$ could have various orthographic realisations, such as *Middellandse Sese lande*, *Middellandse-Sese lande*, *Middellandsesese lande*, etc. ('Mediterranean countries').

Wallis et al. (2012) argues that "[m]any of the research questions we typically wish a corpus to answer can be formulated in terms of variables representing a linguistic choice made by speakers or writers". Against this background they frame linguistic variation as a model of choice: "studies of language variation and change should be primarily conceived as questions of choice" (Wallis et al. 2012:1). Since corpus linguistic analyses are performed *ex post facto* (in contrast to analyses of, for example, experimental data), one needs to account for counterfactuals, i.e. all the possibilities that were available to the author at the moment of writing.

With regard to MWP+N compounds two dependent orthographical variables come into play:

- Letter case variable: The choice between capital and noncapital letters (e.g. ***M****iddellandse* ***S****ee****g****ebied* vs. ***M****iddellandse****s****ee****g****ebied*); and

- Con-/disjunctive variable: The choice between white spaces (*Middellandse See Gebied*), hyphens (*Middellandse-See-gebied*), and no white spaces (*Middellandseseegebied*).

In the case of [[*Middellandse See*]$_{MWP}$ [*gebied*]$_{N}$]$_{N}$ we can therefore postulate a model of choice consisting of 72 opportunities: two letter case variables can occur in three places (2x2x2), and three con-/disjunctive variables can occur in two places (3x3). This model of choice can be represented as an 8x9 contingency table, so that every point of choice is free to vary, i.e. "a genuine choice exists and all cases could theoretically be of one type or the other" (Wallis et al. 2012:4). We can then formalise the research task as one of independent mutual substitution (Wallis et al. 2012:3):

> *Given a corpus, identify all events A that alternate with events {B, C, D, … BT} such that A is mutually replaceable by {B, C, D, … BT} without altering the meaning of the text.*

For each point of choice we can then count occurrences in a given corpus, and determine the conditional probability as:

$$p(\mathrm{A} \,|\, \{\mathrm{A}, \mathrm{B}, \mathrm{C} \ldots \mathrm{BT}\}) = F(\mathrm{A}) / F(\{\mathrm{A}, \mathrm{B}, \mathrm{C} \ldots \mathrm{BT}\}) ,$$

where *F*(A) is the total number of cases (unnormalised frequency) of event A, etc. (Wallis et al. 2012:4; Baayen 2003).

In this workshop presentation we will:

- briefly discuss this model of choice, and the opportunities it affords in operationalising corpus research in the morphographemics space;
- preliminary explore a select few cases of nominal compounds, adjectives, and verbs based on MWPs consisting of two parts (*Middellandse See*), as they appear in three different corpus sources; and
- identify research questions, variables, hypotheses, and baselines for a more comprehensive investigation of these constructions.

References

Baayen, R.H., 2003. Probabilistic approaches to morphology. In R. Bod, J. Hay, & S. Jannedy, eds. *Probabilistic Linguistics*. Cambridge: MIT Press, pp. 229–287. Available at: http://www.sfs.uni-tuebingen.de/~hbaayen/publications/Baayen2003.pdf.

Black, A.W. et al., 1987. Formalisms for morphographemic description. In *Proceedings of the Third Conference on European Chapter of the Association for Computational Linguistics*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 11–18. Available at: http://dx.doi.org/10.3115/976858.976861.

Wallis, S., Bowie, J. & Aarts, B., 2012. That vexed problem of choice. Some reflections on experimental design and statistics with corpora. In *ICAME 33*. Leuven: Catholic University of Leuven. Available at: http://www.ucl.ac.uk/english-usage/staff/sean/resources/vexedchoice.pdf.