Neology and Nomenclature
Concept clarification as a precursor to computational lexicography

Peter Juel HENRICHSEN (pjh@dsn.dk)
Danish Language Council

The Danish Language Council (DSN) is currently developing new concepts and procedures for elicitation of neologisms, with a keen focus on computability. We present some of our newest developments for sharing and discussion.

DSN is responsible for maintaining Retskrivningsordbogen (RO), the dictionary defining the Danish orthographic norm (Schack 2012). RO develops rather slowly, adopting a few hundred new lexemes each year at most, in contrast to the enormous daily production of neologisms in the media and everywhere else. This calls for effective selection criteria. DSN have traditionally used qualitative judgments for verification (Jensen et al. 2014, Jensen et al. 2024), but today's intense media stream has made computational methods necessary. Such methods require formally tightened definitions of certain concepts. In the following, we briefly introduce a few central terms.

An 'exologism' is, basically, a word form appearing in a corpus *C*, though absent in a dictionary *D*. Thus, each pair of language resources, for example,

C_G = DAGW (Danish GigaWord Corpus, Derczynski et al. 2021)
D_R = Retskrivningsordbogen (63k lemmas),

generates a set of candidate tokens. For (C_G,D_R) the characteristic set includes:

'toogfyrre' (*forty-two*)
'Pretoria' (*Pretoria*)
'kiwitærte' (*kiwi-tart*)
'øjebæ' ('øje'+'bæ', *eye+poo*, ≈*ugly-building*)

Some of these tokens are not really foreign to D_R, merely absent for reasons of parsimony; numerals (e.g. 'toogfyrre') and proper names ('Pretoria') are examples of domains deliberately restricted in D_R. We define an exologism as a word form *W* appearing in a corpus *C* while unsupported in a dictionary *D* (formal/computational definitions are in the paper) and a neologism as an exologism in (*C″*,*D″*) where *C″* faithfully represents contemporary language production, and *D″* faithfully represents the shared vocabulary at an earlier state in time (e.g. Δ*t*=10Y). Claiming a neologism *N* thus comes with an obligation to (*i*) establish a corpus *C″*, (*ii*) quantify *N* in *C″*, (*iii*) define Δ*t*, (*iv*) compile a dictionary *D″*, and (*v*) prove that *N* is unsupported in *D″*. While these procedures are by no means trivial, most challenges are inherent to neology as such and must be addressed independent of methodology. We find that the formal approach supports division of labour (procedures *i-v*) and recycling of resources (*D″* and *C″*). It also allows a stricter classification of neologisms. Last but not least, it paves the way for software development.

A current example is DSN's application NeoClink (based on CLINK (Henrichsen 2024), a morphological text parser using categorial grammar and type logic). NeoClink is used for unsupervised extraction of neologisms from text streams. Each input token is broken down into material components ('Morphs'), then analysed for morphological function ('Sequent') and semantic relations ('Semantics'). See table 1 for examples; further details are in the paper.

Token	**Morphs**	**Sequent**	**Semantics**	**Class**
"øjebæ"	[øje][][bæ]	`N X\Y/Y N ==> N`	bæ(øje)	DAN|DAN_{derog}
"antiwoke"	[anti][woke]	`X/X A ==> A`	¬(woke)	DAN_{pre}|ENG
"tjak"	[tjak]	`X ==> X`	?tjak	OOV

Table 1. NeoClink lexemes (reduced CLINK templates). OOV=*Out-of-vocabulary*. `X`,`Y`,`Z`=*category variables*. `N`,`A`=*category constants* (in casu noun and adjective).

DSN's daily text feed from Infomedia (www.infomedia.dk, ≈800M tokens/year) and other text sources are screened on a regular basis using NeoClink, the resulting suggestion list evaluated against Nyordslisten (DSN's manually compiled list of neologisms). NeoClink typically scores very high for recall (>0.9), meaning that most hand-picked candidates are also in NeoClink's output; however much lower for precision (0.2-0.4), NeoClink often over-accepting (*a*) exologisms as neologisms and (*b*) lexical redundancies as exologisms (*a*:*b*≈1:3). This profile makes NeoClink useful as a source of supply while the final decision about inclusion in RO of course remains with the responsible editor.

DSN's traditional classification of neologisms (Jarvad 1995:30-83) is mainly example-based and thus hard to implement. NeoClink's template-based analysis provides a computationally feasible alternative (cf. table 1, 'Class').

Apart from supporting DSN's software development, the semi-formal take on neology has also facilitated our communications, not only internally (lexicographers, computational linguists, assistants), but also across institutional boundaries (e.g. in cooperation with university departments). This is not to say that neology has suddenly become easy; but at least some of the problems now have names.

References

Derczynski, L., et al. (2021) The Danish gigaword corpus. In: *proceed. NODALIDA-23*.

Henrichsen, P.J. (2024) Det Centrale Ordregister. Et indeks over det danske sprog. In: *proceed. NFL-16.* Lund Universitet, 113-126.

Jarvad, P. (1995) Nye ord - Hvorfor og hvordan? In: *proceed. Danish seminar on neology 1995.* Copenhagen University.

Jensen, J.N. et al. (red.) (2014) Neologismer. In: *proceed. DSN seminar om nye ord 2*. DSN-Publications.

Jensen, E.S. et al. (2024) New Words in Danish. In: *proceed. NFL-17* (in prep.). Bergen University.

Jensen, E.S. et al. (2023) (red.) *Nyordslisten. Nye ord i dansk 1955 til i dag.* DSN-Publications. https://dsn.dk/ordboeger/nye-ord-i-dansk

Schack, J. et al. (2012) *Retskrivningsordbogen.* DSN-Publications. https://dsn.dk/ordboeger/retskrivningsordbogen